Implementing association mapping and genomic selection using germplasm from Midwest barley breeding programs

A Dissertation
SUBMITTED TO THE FACULTY OF
UNIVERSITY OF MINNESOTA
BY

Vikas Vikram

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Dr. Kevin P Smith

January 2014

Acknowledgements

This thesis work could be completed only with the help of many people with whom I interacted during my stay at St Paul. First and foremost, I acknowledge the guidance, motivation, enthusiasm, and patience provided by my major advisor, Dr. Kevin P Smith, who has helped me during various phases of this research project and my life at the University of Minnesota. I simply could not wish for a friendlier advisor. I express my sincere gratitude to the members of my graduate committee, Drs. James Anderson, Peter Morrell, Brian Steffenson, and Yang Da. I would also like to thank Dr. Richard Horsley, North Dakota State University for allowing me to work with their germplasm and Dr. Aaron Lorenz, University of Nebraska for sharing his expertise in genomic selection project.

Thanks are due to all members of barley lab who provided a motivating environment to work and helped with various aspects of my research. I thank Ed Schiefelbein and Guillermo Velasquez for helping me in fieldwork and Karen Beaubien for assistance with marker data. The support and amity provided by my friends Ahmad Sallam, Stephanie Navarra, Leticia Kumar, Celeste Falcon, Mohsen Mohammadi, Tyler Tiede and Alexandrea Ollhoff are greatly appreciated. I thank all undergraduate students especially Danelle Dykema, Lauren Schmidt, Jeff Brown, and Fernando Cunha for helping me with research.

I would like to thank Morrell lab group for organizing and maintaining an active journal club in our department. I greatly cherish several discussions happened there

which helped me to develop into a scientist with broader base. I also thank Michael Kantar, Brian Seda and Christopher Schaefer for many fruitful discussions.

This study was funded by Barley CAP, USDA-AFRI NIFA, US Wheat and Barley Scab Initiative, Small Grains Initiative and American Malting Barley Association (AMBA). The financial support I received from the Dr. Smith's lab is appreciated at the highest degree. I received additional support from Burle and Louetta Gengenbach Fellowship, Rahr Fellowship, Charles Burnham/T.T. Chang Fellowship and a travel grant from The Microbial and Plant Genomics Institute. I greatly cherish the opportunity and infrastructure provided by the Department of Agronomy and plant breeding and the University of Minnesota.

I thank my mom - Saraswathy, dad - Vikram, brother - Vinod, and in-laws (Appukuttan and Prema) for your encouragement, prayers, and love as I strive towards my dreams. Thank you very much!

Dedication

This thesis is dedicated to my wife Anaswara, and my baby girl Vrinda for their endless love and support.

Given the stringent requirements set by the malting and brewing industries and the complex nature of traits, marker based breeding will become increasingly important for the improvement of malting barley (*Hordeum vulgare* ssp. *vulgare*). Identification and exploitation of useful and novel alleles at quantitative trait loci (QTL) is crucial to improve genetic gains. We identified marker trait associations and assessed the prediction accuracy of genomic selection for agronomic and disease traits using a collection of 768 breeding lines from two closely related breeding programs in the Upper Midwest. Three hundred progeny lines derived from crosses among 14 parents from the two programs were used as a validation mapping panel in association mapping and to assess genomic prediction accuracy. In general, we found that different sets of QTL were segregating in the two breeding programs. The difference in QTL detected could be due to different genes segregating in the two programs, but could also be affected by differences in marker allele frequencies and linkage disequilibrium between adjacent markers. The genomic prediction accuracies of progeny for six traits were moderate to moderately high, indicating that genomic selection could be successfully implemented for agronomic, disease and quality traits with a range of heritabilities. Our results also indicated that the prediction accuracies were better when training and prediction panels were closely related. The information gained from this study will be valuable to design sound and cost-effective breeding strategies for malting barley whose genetic base has become narrower over time.

Table of Contents

List of Tables

Chapter 1

Chapter 2

Appendix

List of Figures

Chapter 1

Chapter 2

Appendix

Chapter 1

Comparative association mapping and validation of agronomic and disease traits using two North American malting barley breeding programs

Comparison of marker-trait associations among breeding programs could inform collaborative breeding and germplasm exchange. Novel alleles segregating in advanced lines from other breeding programs could provide opportunities for breeders to increase genetic diversity in mature breeding programs without relying on unadapted or exotic parents. We investigated marker trait associations for agronomic and disease traits using a collection of 768 breeding lines from University of Minnesota and North Dakota State University barley improvement programs. The agronomic data were obtained from the Barley Coordinated Agricultural Project trials over a 4 year period (2006- 2009) grown at multiple locations in Minnesota and North Dakota, USA. We used 3,072 single nucleotide polymorphisms to detect trait-marker associations. Across all five traits, we detected 19 and 13 QTL in the Minnesota and North Dakota programs, respectively. Only two QTL were detected in both breeding programs. We were able to validate 20% of these QTL using an independent mapping panel comprised of progeny from parents in the original mapping panel. The difference in QTL detection between the two programs could be due to different genes segregating, but could also be affected by differences between the two programs in marker allele frequencies and linkage disequilibrium between markers and QTL. Analyzing lines from individual breeding programs separately will help breeders to identify QTL relevant to their breeding programs and target environments; however, combining data sets from different breeding programs can

increase power to detect QTL. We conclude that the information gained from this study on a range of traits with different genetic architecture will be valuable to design sound and cost-effective breeding strategies to introduce novel genetic diversity into mature barley breeding programs.

Introduction

A major challenge in plant breeding is balancing the need to introduce novel alleles into breeding populations with the need to maintain favorable alleles that have been accumulated through past breeding and selection efforts (Feuillet et al. 2008; Tester and Langridge 2010). Introduction of large effect genes from unadapted sources has been successful for disease resistance and plant stature genes and can be facilitated by marker assisted selection (Doebley et al. 1993; Lin et al. 1996; Young 1996; Hajjar and Hodgkin 2007). In contrast, improving complex agronomic traits has been challenging to breeders as these traits are controlled by many small effect genes (Schon et al. 2004; Bernado 2008).

Mapping quantitative traits has been done extensively using bi-parental populations. Often these studies have employed exotic parents because of the need to generate sufficient phenotypic variation in the population and adequate genome coverage with genetic markers for mapping. Since mapping parents are frequently unrelated to the target breeding germplasm, the information gained from these studies has had a relatively minor impact on crop improvement for quantitative traits (Young 1999; Bernardo 2008).

More recently, association mapping approaches have been applied to panels of elite breeding germplasm to identify QTL and alleles that may be more directly useful in breeding. These studies have identified QTL for many traits including disease resistance, heading date, kernel size and grain quality (Breseghello and Sorrells 2006; Massman et al. 2010; Wang et al 2011; Zhou 2013). All of these studies either used breeding lines from a program or a diverse set of lines from multiple breeding programs to identify

markers associated with the traits of interest. Exploiting mapping panels derived from breeding programs using association mapping provides an opportunity to leverage multi-year and multi-location phenotypic data to characterize relevant alleles with little or no additional cost (Rafalski 2002). Coupling this approach with increasingly less expensive genotyping opens the door to a relatively untapped resource for genetic study. Designing mapping panels from different breeding programs creates the possibility of comparative analysis of QTLs from populations with different breeding histories. Understanding the genetic architecture of traits in different breeding programs should inform strategies to employ markers in breeding for quantitative traits.

Although association mapping is convenient for breeders to identify QTL, the probability of making type I or type II error is higher than traditional linkage mapping (Pritchard et al. 2000), and follow-up validation studies are necessary before undertaking any marker-assisted selection (Lander and Kruglyak 1995). Cryptic relationships among lines within a breeding program and multiple testing could increase type I error rate or false positive associations. Mixed model analysis with appropriate individual relatedness matrices and stringent statistical correction for multiple testing can be employed to reduce the detection of false positive associations (Yu et al. 2005; Carlson et al. 2004). However, there is a tradeoff between type I and type II error and attempts to reduce one will unfortunately increase the other (Belknap et al. 1996). QTL can be validated by comparing near isogenic lines (NILs) with contrasting alleles for the QTL of interest (Van Berloo et al. 2001; Pumphery et al 2007; Navara et al. *in review*). Because development and testing of NILs in multiple genetic backgrounds and environments is

very resource intensive, often allelic effects are tested only in a single genetic background. Alternatively, QTL identified in a population could be validated using independent mapping populations constructed from closely related genotypes from the primary mapping study (e.g., Canci et al. 2004). Using a mapping panel constructed with progeny derived from multiple parents would permit testing of QTL effects in many genetic backgrounds.

The challenge of effective introgression of novel genetic diversity in mature breeding populations is exemplified in the case of U.S. malting barley (*Hordeum vulgare* ssp. *vulgare*). Industry acceptance of new varieties requires that they perform similarly in malting and brewing to existing approved cultivars in addition to having superior agronomic performance and disease resistance (Horsley et al. 1995). Because measuring malting quality is very costly and time consuming, barley breeders in the upper Midwest U.S. have relied on advanced cycle breeding where superior parents within a single germplasm pool are crossed among each other to accumulate favorable alleles. Alternative attempts to improve yield, protein and kernel discoloration by introducing parents from genetically diverse breeding programs required three to six breeding cycles to achieve modest gains (Gebhardat et al. 1992; Goblirsch et al. 1996; Peel and Rasmusson 2000). Thus, advance cycle breeding has been the primary approach to achieve steady gains in quantitative traits and the release of successful varieties; however, it has also led to a reduction in genetic diversity (Rasmusson and Phillips 1997; Condon et al. 2008a, 2008b).

The University of Minnesota (MN) and North Dakota State University (ND) barley breeding programs are ideal to explore the potential of collaborative breeding because both have similar breeding objectives, geographic range, and disease pressures. In addition, both followed advanced cycle breeding from early 1950's, with limited exchange of germplasm between the programs. Advanced cycle breeding approaches have been shown to reduce genetic diversity in the breeding programs over time and fix favorable alleles for important traits that are under selection (Condon et al. 2008a, 2008b). If the different breeding histories of the MN and ND breeding programs resulted in selection for different favorable alleles, then these alleles could complement each other in a collaborative breeding effort. Novel alleles segregating in elite backgrounds from other closely related breeding programs could reduce the linkage drag often associated with wide crosses with parents outside of a breeding population.

In this study, we utilize association mapping to characterize and contrast two advanced cycle breeding programs for agronomic and disease traits. Primary breeding objectives for the MN and ND breeding programs include increasing yield and resistance to FHB and maintaining favorable malting quality. Our specific objectives were to: (1) Map QTL associated with Fusarium head blight severity, deoxynivalenol concentration, grain yield, plant height and heading date using breeding lines from two North American barley improvement programs, (2) Determine the consistency of QTL locations between breeding programs, (3) Determine whether combining breeding data sets will enhance power to map QTL, (4) Validate the identified QTL using an independent population comprised of progeny from crosses among parents in the original mapping panel.

Materials and methods

Plant materials and experimental design

We studied plant height, heading date, grain yield, FHB severity and DON accumulation in six-row spring malting barley germplasm from the University of Minnesota (MN) and North Dakota State University (ND) barley breeding programs. The entire population consisted of 768 barley breeding lines (384 lines from each program) that are representative of lines entering first year yield trials. Prior to entry into first year yield trials, these breeding lines qwere evaluated in single row plots (head rows) and selected for acceptable heading date, plant height, FHB resistance, and lodging resistance (see for example Smith et al., 2012). Within a program there were 4 sets of 96 lines referred to as CAP I – CAP IV as part of the Barley Coordinated Agricultural Project (CAP; www.barleyCAP.org). All lines were developed by single seed decent to the F4 or F5 generations and subsequent bulked generations were used in field trial experiments. Agronomic traits (plant height, heading date, grain yield) were evaluated as a part of the normal breeding program trials. Thus, each program evaluated their own lines using randomized complete block designs with two replications carried out at multiple test locations in Minnesota and North Dakota. Different sets of lines were evaluated in different experiments resulting in an unbalanced data set. All the lines were evaluated in at least three and as many as eight experiments. For agronomic traits, the plot sizes were 1.9 m^2 for MN trials and 3.3 m^2 for ND trials. Repeated check cultivars were Baronesse, Harrington, Lacey, Legacy, MNBrite, Robust, Stander, Stellar, and Tradition in the MN trials, and Baronesse, Harrington, and Robust in the ND trials. Evaluation of DON and

FHB were done in collaborative experiments where all of the lines for the two programs were included in a single experiment for a given CAP set (I – IV). The details of the experiment setup and disease evaluation protocols for FHB severity and DON accumulation can be found at Massman et al. (2011). Briefly, entries were planted in single row plots, inoculated with *Fusarium graminearum* as either macroconidia or grain spawn, and mist irrigated following inoculum application. We defined three mapping panels using above 768 barley breeding lines: (1) University of Minnesota panel (M; 384 lines); (2) North Dakota State University panel (N; 384 lines); (3) combined panel (M+N; 768 lines).

Phenotype data

Heading date was measured as the number of days after planting in which at least 50% of the spikes in a plot were emerged at least half way from the boot. Plant height was measured as the length of the plant (cm) from the soil surface to the tip of the spike excluding awns and was recorded as the mean of two measurements per plot. Plots from yield trials were harvested with a Wintersteiger Master Elite combine plot harvester (Wintesteiger, Ried, Germany) and grain yield was recorded as kilograms per hectare at 14 percent moisture content. Disease severity was measured as the percent infected kernels based on visual ratings of 10 arbitrarily selected spikes and mature grain was harvested from plots, ground, and analyzed by gas chromatography and mass spectrometry as described by Mirocha et al. (1998). All phenotypic and genotypic (see below) data for analyses are stored in a publically available database, The Hordeum

Toolbox (THT; www.hordeumtoolbox.org) developed as a part of Barley CAP (Blake et al. 2012).

Genotype data

A detailed description of Barley CAP genotyping procedures was provided by Massman et al. (2010). Briefly, DNA was extracted from a single plant from the F4 or F5 bulk seed source and two barley oligonucleotide pool assays (BOPA; Close et al.2009) each containing 1,536 allele specific SNP markers were used to genotype the breeding lines. We screened all markers for missing information, minor allele frequency and excessive heterozygosity to eliminate potential genotyping errors. Markers with greater than 25% missing information, minor allele frequency lesser than 0.001, or excess heterozygosity in M+N data set were excluded from our analyses. Out of 3,072 SNP markers screened, 2,110 met the above criteria and were used for further analysis.

Linkage disequilibrium and population stratification

We calculated linkage disequilibrium (LD) between markers using HAPLOVIEW v.4 (Barrett et al. 2005). When multiple markers had the same genetic map location, the marker with the least missing information was retained for LD calculations. We calculated pairwise measures of LD as r^2 derived from two-marker expected maximum-likelihood values and used this data to characterize LD decay over genetic distance in all three mapping panels. Because QTL detection power is a function of LD between marker and QTL, we also calculated LD between adjacent markers for the M and N mapping panels.

We used principal component scores of breeding lines estimated from genome wide SNP markers to determine sub population grouping. First two principal components were plotted against each other to study population structure. We also calculated F_{ST} coefficient (fixation coefficient) to describe genetic differences between M and N mapping panels and differentiation in allele frequencies between the two populations (Weir & Cockerham 1984).

To model family structure and cryptic relatedness among breeding lines, we estimated identity by descent based kinship matrices among lines using genome wide SNP markers. We estimated kinship in all three mapping panels using the additive model implemented in the EMMA package in R (Kang et al. 2008). For association mapping in plant breeding programs, it has been previously shown that correction with the kinship matrix alone is sufficient (Bradbury et al. 2011; Wang et al. 2011; Wu¨rschumet al. 2011a, b). We generated a Q matrix using first three principal components. The results of association mapping using models with PCA based Q and kinship model (Q+K) and the kinship (K) alone model were very similar (data not shown). Therefore, all the results reported are based on the K model.

Statistical analyses

We checked traits for obvious outliers based on bivariate plots of raw data. The residuals were checked for normality (via different normality tests using histogram, box plot, stem-and-leaf plots and QQ plot) and homoskedasticity (via residual plots) using PROC UNIVARIATE (SAS®, version 9.3.1). Observations with a studentized residual absolute value greater than three were removed from our analysis as outliers.

Because our phenotype data were collected from many trials that spanned over four years and two breeding programs, we employed a two step analysis to calculate best linear unbiased estimates (BLUEs) for each line. First, common repeated check varieties from all trials were used to estimate respective trial effects for all traits using the model $\mathbf{Y} = \mathbf{X}\alpha + \mathbf{C}\beta + e$; where **Y** is the vector of check phenotypes, **X** is an incidence matrix relating trial effects to **Y**; **α** is a vector of trial fixed effects to be estimated; **C** is an incidence matrix relating check line information to **Y**; β is a vector of check effects; and e is a vector of residual variation where $e \sim \mathbf{N}(0, \sigma^2_e \mathbf{I})$. Second, each trait in each trial was adjusted for respective trial effects, and best linear unbiased predictions (BLUEs) of individual lines were calculated using the model $\mathbf{Y_{adj}} = \mathbf{L}\gamma + e$; where $\mathbf{Y_{adj}}$ is the vector of trait adjusted for corresponding trial effects. **L** is an incidence matrix relating line information to $\mathbf{Y_{adj}}$; γ is a vector of fixed effect of breeding lines to be estimated; and e is a vector of residual variation where $e \sim \mathbf{N}(0, \sigma^2_e \mathbf{I})$. QTL mapping was then performed on best BLUEs of individual lines.

We used PROC MEANS to calculate trait means, and PROC MIXED to estimate variances (SAS®, version 9.3.1) from Y_{adj}. Broad sense heritability (H) on an entry means basis was calculated from these variances using the equation $H = \sigma^2_L / [\sigma^2_L + \sigma^2_e/r]$; where r is the number of trials a line was tested.

Association mapping:

We performed association mapping analysis using a mixed model approach that corrected for kinship implemented in R package EMMA (Kang et al. 2008) with the model $\mathbf{Y} = \mathbf{X}\alpha + \mathbf{u} + e$; where, **Y** is the vector of BLUE of individual lines; **X** is a matrix of fixed effects

of SNP marker information; α is a vector of SNP effects to be estimated; **u** is random polygene background effects where $\mathbf{u} \sim N(0, \sigma^2_g \mathbf{K})$. σ^2_g is genetic variance and **K** is the kinship matrix as described above and e is a vector of residual variation where $e \sim N(0, \sigma^2_e \mathbf{I})$. σ^2_e is the error variance and I was the identity matrix. The phenotypic variance of each significant SNP was estimated using PROC GLM by fitting the model $\mathbf{Y} = \mu\mathbf{M} + e$, where M is the vector of SNP genotype and μ the SNP effect. A FDR cutoff of 1% was used to declare a QTL (Benjamini and Hochberg 1995).

We investigated the effect of population size and mapping panel composition on the number of QTL identified using a random re-sampling analyses. Association mapping was repeated on ten random re-sampled subpopulations (M+N: 384 lines) generated with replacement from our combined dataset (M+N) with 768 lines. All re-sampled data sets consisted of lines from M and N panels in 1:1 ratio. The number of SNP detected with a – log p above 3 in M+N:384 resampled subpopulation was compared with M+N.

QTL validation

We subjected the QTL identified in the M, N and M+N mapping panels to validation using a set of progeny (n = 300) developed from crosses among parents that were members of the M and N mapping panels. Each progeny line in the validation panel was advanced to the F3 by single seed descent and tested in 2011 and 2012 using $F_{3:5}$ seed using an augmented block design. Agronomic traits were evaluated at St Paul, MN, Crookston, MN, and Nesson Valley, ND in 2011 and in St Paul, MN, Crookston, MN, Nesson Valley, ND and Fargo, ND in 2012. For FHB trials, we followed the same augmented block design except that the trial was replicated twice at each location.

Disease traits were evaluated in St Paul, MN, Crookston, MN, Osnabrock, ND and Langdon, ND in 2011 and in St Paul, MN, Crookston, MN, Osnabrock, ND and Langdon, ND in 2012.

A detailed description of the population used for QTL validation is provided in chapter 2. Briefly, the population consisted of 100 progeny lines from each of three different cross types: (1) crosses among parents within the MN breeding program (MxM), (2) crosses among parents within the breeding ND program (NxN), and (3) crosses using parents from both breeding (MxN). The phenotype data was collected as described above.

Because the progeny panel was originally developed as part of ongoing genomic selection project in our lab (see chapter 2.), we developed an Illumina VeraCode Custom OPA containing 384 selected SNP markers (i.e., ~1 marker/3 cM), and obtained genotype data using Illumina's BeadXpress instrument. These markers were selected from barley OPA1 and 2 based on polymorphic information content among the 14 parents and even distribution across genome (Close et al. 2009). The markers were screened for quality using the parameters described above and a final set of 340 markers were used to perform association mapping using the procedures described above.

Results

Phenotypic and genetic variation

Means, ranges, and heritabilities for plant height, heading date, grain yield, FHB and DON from M and N panels are shown in Table 1. Trait heritabilities were similar in both

panels except for heading date, where the M panel had lower additive genetic variance than the N panel.

Our principal component analysis indicated that the first two principal components of SNP marker data explained 10.9% and 5.8% of the variation. The first principal component separated the breeding lines into respective breeding programs, and second principal component explained the variance among lines within each breeding program (Fig. 1). The per locus Fst ranged from zero to 0.81 across the genome (Fig. S1). The median pair wise kinship was 0.87 in M and 0.84 in N mapping panels (Fig. S2).

Linkage disequilibrium and allele frequency

The average adjacent pair-wise LD was higher in the M panel than ($r^2 = 0.55$) than N panel ($r^2 = 0.30$). The LD decay curve indicated that LD extended to 7 cM in both breeding programs before it decayed to background level (Fig. 2). LD decayed faster in N compared to M.

Because QTL detection power is a function of allele frequency and LD, we characterized the differences between the M and N mapping panels for minor allele frequency and adjacent marker LD (Fig. 3). Differences in allele frequency and adjacent marker LD varied widely along all seven chromosomes particularly at QTL regions for MN and ND mapping panels. We observed three notable large regions of LD difference: 1H: 30-45cM, 2H:95-115 cM, and 7H: 80-100 cM. Chromosome 6H displayed minimal difference in adjacent marker LD.

Markers trait associations

Genome wide scans of M, N, and M+N mapping panels detected 39 marker-trait associations (Table 2; Fig.4). The number of QTL detected for each trait ranged from 3 to 17: plant height (8), grain yield (7), heading date (17), DON (4), and FHB (3). We observed three QTL regions that harbored more than one QTL: CH4H:24-26 cM, CH6H:0-3 cM, and CH7H:86.44 cM. The CH4H:24-26 cM region was associated with grain yield, FHB and plant height (Fig. 4). Interestingly, 62% QTL (24 of the 39) were detected in either M or N mapping panel. Only two QTL were detected in both M and N: heading date locus at 5H:29.9-33.09 cM and 7H:39-45 cM. Furthermore, a comparison of SNP p-values for five traits indicated that most SNP markers detected at a significance of -logP greater than 2 were breeding program specific and only nine SNP markers were similarly significant in both breeding programs (Fig. 5). In contrast, 195 SNP markers had -logP greater than 2 in M panel, but less than 2 in the N panel; and likewise, 205 SNP markers had -logP greater than 2 in N panel, but less than 2 in the M panel. For DON and grain yield, no QTL were detected in ND mapping panel that exceeded the FDR cutoff of 1%.

On average, the combined M+N mapping panel detected more QTL than individual mapping panels (either M or N). Of the 17 heading date QTL identified, 7 QTL were detected both in the individual panels and M+N, 9 QTL detected in individual panels were not detected in M+N, and 1 QTL was detected in M and N but not in M+N, and 3 QTL detected in M+N were not detected in M or N. For plant height, of the total 8 QTL identified, 5 QTL were detected both in individual panels and M+N, all QTL

detected in individual panels were also detected in M+N, and 3 QTL detected in M+N were not detected in M or N. For grain yield, of the total 7 QTL identified, 2 QTL were detected both in individual panels and M+N, only 1 QTL detected in an individual panel was not detected in M+N, and 4 QTL detected in M+N were not detected in either M or N. For FHB, of the 3 QTL identified, 1 QTL was detected in individual panels and M+N, 2 QTL detected in individual panels were not detected in M+N, and all QTL detected in M+N were also detected in M or N. For DON, of the 4 QTL identified, 1 QTL were detected in individual panels and M+N, 2 QTL detected in individual panels were not detected in M+N, and 1 QTL detected in M+N were not detected in M or N. Comparisons between the combined (M+N) and individual (M, N) programs must take into account the differences in population size. However, when population size was held constant, the mean number of SNP detected above logP threshold of three in the combined mapping panel with 384 lines (M+N:384) and M or N mapping panels were same (Fig. 6).

QTL validation using an independent mapping panel

To independently validate the QTL detected in the M, N, and M+N panels, we conducted association mapping for all traits using an independent progeny panel tested in 4-6 environments (Table S3). For the five traits, we validated 8 out of 39 (~20%) QTL that were previously identified using M, N, and M+N mapping panels (Table2 and Fig. S3-7). We validated one QTL for DON (2H:81.33 cM and 7H:39.04 cM), two for FHB (4H:24cM, 3:133.14 -145.89 cM), two for grain yield (4H:24cM; 6H:9.06-12.54cM), two for plant height (4H:24.59cM; 6H:56.48cM) and one for heading date (7H:39.04 -

43.38cM). Interestingly, the plant height QTL at 6H:56.48 cM (12_30144) and heading date QTL at 7H:39.04 cM (12_10218) were validated with same SNP in both the original mapping panel and the validation panel. The plant height SNP 12_30144 was highly significant in M, M+N and validation mapping panels. In all other cases, we validated the genomic region with a different nearby SNP (within 5 cM). Because our validation panel was genotyped with a smaller subset of the original marker set, it often did not contain the most significant SNP associated with traits in M or N mapping panel. The locus 4H:24-26 cM was found to be associated with multiple traits (FHB, plant height, and grain yield) in both the original mapping panels and validation panel. As expected, the total number of QTL detected in our validation data set in general was lower than M, N or M+N.

Discussion

A comparative analysis of two breeding programs with similar objectives and geographic regions surprisingly revealed that nearly all QTL were specific to one breeding program or the other. Earlier studies reported marker trait association on heading date, FHB severity and DON using subsets of germplasm used in this study (Massman et al. 2010; Wang et al. 2011). However, those studies were based on combined analysis of breeding lines across multiple breeding programs. The differences in QTL detection between the two programs could be due to different genes segregating in the two programs, but could also be affected by differences in marker allele frequencies and patterns of LD. We validated our mapping study using an independent progeny panel tested in new environments. The insights into the genetics of these traits will be valuable to design

sound and cost-effective breeding strategies to introduce novel genetic diversity to elite malting barley whose genetic base has become narrower over time.

Marker trait association using breeding germplasm

We detected both novel and previously reported QTL for all five traits distributed among all seven chromosomes. Notably, we detected association signals near previously reported QTL for heading date 2H:58 cM and 7H:39 cM (Mesfin et al. 2003; Rostoks et al. 2005; Horsley et al. 2006; Nduulu et al. 2007; Wang et al. 2011); plant height 3H:52 cM, and 3H:130 cM (Pillen et al. 2003); grain yield 7H:99-109 cM (Pillen et al. 2003; Comadran et al. 2008); and FHB 4H:24-36 cM, 3H:145 cM (Mesfin et al. 2003). Consistent with a previous study that used a subset of data set used in our study, we detected associations for DON at 5H:190-192 cM and 6H:42-67 cM (Massman et al. 2012). Although 2H:81 cM has been reported as a FHB QTL in a Fredrickson/Stander mapping population (Mesfin et al. 2003), we found this locus to be associated with DON and not FHB. Interestingly, the SNP marker (12_10859) associated with this locus had high –log p value (>3.7) in M+N and N and the same region was found to be associated with DON in the validation panel (see below for details). Several novel QTL were detected and independently validated in our study: HT 6H:56.48 cM; YD 4H:26.19 cM; YD 6H:9.06 cM; DON 2H:81.33 cM and HD 7H:86.44 cM. Our ability to detect new and previously reported loci indicates that breeders could potentially utilize association mapping as a tool to leverage vast amounts of data collected over years and locations to identify relevant marker trait associations without conducting any new phenotyping.

Previous studies on spring barley indicated strong relationships between FHB severity and DON concentration with heading date and plant height (de la Pena et al. 1999; Mesfin et al. 2003; Dahleen et al. 2003; Horsley et al. 2006; Massman et al. 2010). A subsequent large biparental mapping study using 1,500 recombinant near isogenic lines characterized the bin 8 region of Chromosome 2H and reported that that the association between FHB and heading date was due to tight linkage rather than pleiotropy (Nduulu et al. 2007). We identified six coincident QTL for barley scab in our study, out of which 4 were associated with heading date and 2 were associated with plant height.

The CH4H:24-26 cM was associated with multiple traits including grain yield, FHB and plant height (Fig. 4). Previous studied including a detailed resequencing study indicated that this region contained *INTERMEDIUM* (*INT-C*) genes that influence row type morphology, spikelet size and fertility (Lundqvist and Lundqvist 1989; Ramsay et al. 2011; Druka et al. 2011). Furthermore, the infertile lateral floret locus (int-c) was also associated with reduced FHB and DON accumulation (Zhu et al. 1999). Mesfin et al. (2003) indicated that two-row barley lines had lower FHB infection and DON accumulation than six-row barley. We also observed co-localization of FHB and yield QTL with *INT-C* genes, a result that supported many previous studies. Recombinants at this region that possess favorable alleles for both traits should be used as parents for future breeding and selection.

QTL validation using an independent progeny panel derived from advanced breeding lines

Despite the power, resolution and versatility of association mapping studies, the ability to control for false positive associations arising from cryptic relationships among lines and multiple testing is often insufficient. This necessitates follow-up validation studies using an independent mapping panel preferably tested in multiple new environments (Pumphrey et al. 2007). An alternative approach is to develop NILs (Navara and Smith *in review*). In the current study, we validated the QTL identified using an independent mapping panel comprised of progeny developed from crosses among breeding lines from MN and ND breeding germplasm.

We were able to detect 20% of the QTL identified in our original mapping study in our validation panel. In general, the QTL that were validated had strong association signals in the original mapping panels. The validated SNP 12_30144 at plant height QTL 6H:56.48 cM had largest –log P value (9.5 and 7.0) signal in N and M+N mapping panel. Similarly, the heading date SNP 12_10218 we validated at QTL 7H:39.04 cM also had largest –log P value signal in M. For all other validated QTL regions, a nearby SNP had large –logP value signals in M, N or M+N mapping panels. Our ability to validate markers associated with traits using an independent population indicates that breeders could use GWAS to identify reliable markers-trait associations in breeding germplasm. In our case, we found the direction of allelic effects of all validated markers remained same across populations and environments. Furthermore, because both QTL detection and QTL validation were performed using elite breeding lines relevant to breeders, breeders could

readily use these markers and breeding lines for marker assisted selection and improvement.

Because many small effect genes only explained a small percentage of total genetic variance, the repeatability of GWAS in general was reported to be low (Yang et al. 2010; Pritchard and Cox 2002). Our results also supported these findings. Alternatively, both power and resolution of QTL mapping could be improved using specially designed synthetic populations (King et al. 2012).

There are at least four possible reasons why we were only able to validate 20% of the QTL detected in M, N and M+N mapping panels. First, because our validation panel was genotyped with a lower density marker panel (i.e., ~1 marker/3 cM), SNPs that were strongly associated with QTL in M or N mapping panel might be missing from our validation panel. For example, of the 8 QTL validated in our study, our validation panel was genotyped with the same significant SNP for only two of the QTL. We detected those two QTL with same SNP, and the remaining QTL with a nearby (within 5 cM) SNP. Although marker density was low, it must be noted that all 340 SNP used for validation were adequately distributed on all seven chromosomes (Table S3). Second, if the MAF of a SNP is lower (or as in extreme case if a SNP fixed) in validation panel compared to M, N or M+N, the SNP will remain undetected. A QTL with a small effect is expected to suffer more from low allele frequency than a large effect QTL. Third, the QTL detection power of validation panel will be lower than M, N, or M+N panels due to its smaller (300) population size (Long and Langley 1999). Finally, stringent correction for multiple testing and population structure can also reduce repeatability of GWAS

(Carlson et al. 2004; King et al. 2012). In spite of these limitations, our ability to replicate some QTL suggests MAS may be effective in breeding.

Breeding programs differed in marker trait associations

Because the MN and ND programs had similar breeding objectives, geographic target areas and disease pressures, we were interested to see whether a joint association analysis could provide insight that would facilitate collaborative breeding. For all five traits included in our study, there were virtually no QTL in common between the MN and ND breeding programs. This result was rather unexpected given the nature and goals of the breeding programs. Several possibilities could explain this lack of congruence. First, barley breeding and improvement practices in the MN and ND breeding programs dating back to early 1900's (Horsley et al., 1995) and differences in selection pressure may have led to significant differentiation in germplasm. Good by good crossing, a classical feature of advanced cycle breeding, was necessary to create favorable gene combinations and satisfy strict industry requirements for malting barley (Peel and Rasmussen 2000; Rasmussen and Phillips 1997). In the MN program, the practice of advanced cycle breeding resulted in gradual differentiation of the germplasm over time that was accompanied by a reduction of genetic variance that was more apparent at specific regions in the genome (Condon et al. 2008a). Even if the same alleles are under positive selection, differences in the intensity of selection or even drift could result in allele frequency differences that effect QTL detection. This was likely the case for the marker with the strongest signal for heading date at 2H: 58-65 cM that was identified in mapping panel N (MAF= 0.02) but not in M (MAF=0.002).

The likelihood of detecting a QTL with a marker increases as the LD between the marker and QTL increases (Long and Langley 1999; Zhu, 2010). Differences in adjacent marker LD between mapping populations might explain differences in QTL detection between the MN and ND programs. In general, we observed large variation in adjacent marker LD along all seven chromosomes. In some regions of the genome this pattern was very different between the two breeding programs. For example, markers at MN heading date QTL at chromosome 1H: 34 cM exhibited stronger adjacent marker LD in the M mapping panel ($r^2 = 0.80$) compared to the N panel ($r^2 = 0.20$). Even large effect QTL will not be detected unless there is sufficient LD between the marker and QTL.

In the early nineties, the MN and ND breeding programs initiated breeding efforts to improve FHB resistance. To avoid duplication of effort, both programs used different sources resistance. The prominent sources used by the MN program were Frederickson, Zhedar1, Comp351, Hor211 and Atahualpa, whereas the ND program used Zhedar2, CI4196 and Clho6611 (Huang et al. 2012). These parents represent genetically diverse sources of resistance for FHB. Independent breeding with these sources could have resulted in different resistance genes segregating within the MN and ND programs. Previous diversity analysis also indicated that FHB resistant sources had broad genetic diversity with haplotypes from Chevron, Fredrickson, and CIho4196 being most prevalent in cultivated barley (Huang et al. 2012). Tracing the marker haplotypes at FHB and DON QTL back through the pedigrees of these breeding lines to their corresponding donor parents could help confirm whether the programs are exploiting different resistance genes.

The fact that two closely related breeding programs did not share common QTL suggests that breeders should not expect more distantly related breeding programs to share many common QTL either. Markers discovered through association mapping in one breeding program may not be directly useful for MAS when applied within another. However, when mapping studies of multiple breeding programs reveal diverse QTL, there should be potential to improve traits by making crosses between programs. In such crosses, markers associated with QTL identified in the different programs could be used to pyramid QTL and exploit potentially complementary genetic diversity for traits of interest.

Pooling datasets from breeding programs improved the power of QTL detection

The power of association mapping is largely a function of amount of variance explained by causal loci, allele frequency, and LD between markers and causal loci (Jannink and Walsh 2002). In addition, mapping panels with larger population sizes and genetic variances will also increase the power to detect QTL (Bevis 1998; Long and Langley 1999; Schon et al. 2004). Therefore, it is appealing for breeders to pool their data sets, particularly for traits that are expensive to phenotype (e.g., DON concentration or malt quality). As expected, our combined M+N mapping panel detected more QTL than either the M or N mapping panels alone. To determine whether the increased power was due to larger population size or increased genetic variance, we conducted a resampling analyses that held population size constant. When this was done, the number of SNP above -logp threshold of three identified in M+N:384, M, and N panels was similar (Fig. 6).

Therefore, in our case, the larger power of M+N mapping panel was a result of sample size rather than genetic variance.

Our combined data set did not identify all QTL identified in either MN or ND germplasm; although, we did detect many overlapping QTL (Table 2). For example, the height QTL at 3H:52 cM detected in mapping panel M was not observed in M+N. Interestingly, the allele frequency of the most significant SNP in that region was ~0.11 in M but was fixed in N. When M and N panels we combined, the allele frequency of the SNP was 0.05. The low MAF of the SNP in M+N panel might be one reason why it did not have enough power to detect the QTL even though the M+N mapping panel had twice the sample size. Therefore, depending on trait architecture, combined datasets might limit breeder's ability to identify all relevant QTL in their germplasm.

Implications for breeding

Using GWAS in elite breeding germplasm provides new opportunities for breeders to characterize and exploit genetic diversity in breeding. Some of the limitations of GWAS in elite breeding germplasm include longer blocks of LD leading to lower mapping resolution, small portions of overall genetic variance explained by identified causal loci, and low power in detecting QTL using a population already subjected to strong selection (Massman et al. 2010). Comparative GWAS between breeding programs can provide useful insight for collaborative breeding and germplasm exchange. This provides opportunities for breeders to incorporate novel and useful alleles to increase genetic diversity without relying on unadapted or exotic parents. This is particularly important as

breeding programs mature, differentiate from founding parents, and accumulate favorable allele complexes.

Because the identification of QTL in GWAS was largely based on a statistical significance cutoff and identified QTL can only explain a small fraction of variance for a trait, selection and breeding based on those QTL alone may have limited utility. For complex traits controlled by large number of small effect QTL (e.g., grain yield), a better strategy will be one that can capitalize on all the biologically important QTL without imposing a statistical cutoff. Simple backcrossing or recurrent selection could be used to introgress large effect genes. However, as the number of QTL increases the number of progeny required to identify lines fixed for all favorable alleles quickly becomes impractical. Marker based prediction models without QTL mapping as implemented in genomic selection might be a better option for the simultaneous improvement of polygenic traits that fit infinitesimal model (Meuwissen et al 2001; Bernado 2008).

Table 1. Trait descriptive statistics for the Minnesota (M), North Dakota (N), Minnesota and North Dakota combined (M+N), and validation panels used for association mapping.

Trait	n[a]	Mean	Minimum	Maximum	H
Plant height (cm)					
M	384	79	57	106	0.85
N	378	80	48	115	0.88
M+N	768	79	48	115	0.87
Validation	300	80	56	105	0.73
Grain yield (Kg ha^{-1})					
M	384	5,875	2356	10,165	0.48
N	378	4,754	883	9,410	0.55
M+N	768	5,207	883	10,165	0.58
Validation	300	4,929	2,637	8,394	0.45
Heading date (days)					
M	384	56	48	68	0.78
N	378	58	48	71	0.94
M+N	762	57	48	71	0.90
Validation	300	56	48	69	0.82
FHB (severity %)					
M	384	16.6	0.2	92.4	0.48
N	384	23.2	0.1	87.8	0.57
M+N	768	19.7	0.1	92.0	0.56
Validation	300	7.4	0.4	30.0	0.45
DON (ppm)					
M	384	24.2	0.8	129.4	0.49
N	384	27.8	0.9	112.3	0.46
M+N	768	25.8	0.8	129.4	0.48
Validation	300	13.0	0.07	43.6	0.57

[a]n is the number of lines tested, Mean is the arithmetic mean, Minimum is the minimum trait value, Maximum is the maximum trait value, and H is the broad sense heritability.

Table 2. Summary of QTL identified by association mapping in six-row spring barley using the Minnesota (M), North Dakota (N), and combined (M+N) mapping panels.

QTL[a]	Mapping panel	Frequency of 'A' allele[b]			Allelic effect[c]			QTL validation[d]
		M	N	M+N	M	N	M+N	
Plant height								
HT1H: 65.53	M+N	0.02	0.02	0.02	–	–	1.9	
HT2H: 116.49	M+N	0.96	0.98	0.97	–	–	-3.9	
HT3H: 52 – 56	M, M+N	0.89	1.00	0.95	6.7	–	6.5	
HT3H: 130	M+N	0.98	0.98	0.98	–	–	-3.1	
HT4H: 26.19	M, M+N	0.36	0.01	0.19	4.0	–	3.0	**X**
HT6H: 0	N, M+N	0.94	0.84	0.89	–	4.9	3.9	
HT6H: 42 – 43	N, M+N	0.09	0.16	0.13	–	-4.9	-0.6	
HT6H: 56.48	N, M+N	0.01	0.17	0.09	_	-4.9	-4.7	**X**
Grain yield								
YD4H: 26.19	M+N	0.37	0.01	0.19	–	–	-119	**X**
YD5H: 2.81	M, M+N	0.82	0.96	0.89	108	–	65	
YD6H: 9.06	M	0.97	0.89	0.93	-265	–	–	**X**
YD7H: 61 – 62	M, M+N	0.96	0.75	0.86	232	–	84	
YD7H: 73.75	M, M+N	0.96	1.00	0.98	294	_	288	
YD7H: 86.44	M, M+N	0.96	1.00	0.98	294	_	288	
YD7H: 91.79	M+N	0.97	1.00	0.97	_	_	-288	
FHB								
FHB3H: 145.89	N	0.09	0.01	0.05		7.1		**X**
FHB4H: 12.02	N	0.40	0.36	0.38		1.9		
FHB4H: 24.59	M, M+N	0.37	0.01	0.19	-2.8		-2.5	**X**
DON								
DON2H:81.33	M+N	0.99	0.68	0.83			-2.28	**X**
DON3H:32.83	M, M+N	0.98	0.98	0.98	-4.89		-3.61	
DON5H:189.6	M	0.30	0.71	0.51	2.37			
DON6H:3.11	M	0.12	0.41	0.26	-2.97			
Heading date								
HD1H: 34.83	M	0.05	0.40	0.22	0.21			

HD1H: 107.55	M	1.00	0.96	0.98	-0.01			
HD2H: 0	N	0.00	0.02	0.01			-2.2	
HD2H: 19.47	N, M+N	0.05	0.01	0.03	1.98		1.59	
HD2H: 33.74	N						1.27	
HD2H: 58-65	N, M+N	1.00	0.98	0.99		-2.32	-1.46	
HD3H: 100.6	M				-0.95			
HD3H: 173.17	M	0.02	0.02	0.02	0.99			
HD4H: 0	N, M+N	1.00	0.98	0.99			1.60	
HD4H: 48-51.3	M, M+N	0.01	0.01	0.01	-2.7		-1.37	
HD4H: 61.04	M	0.16	0.41	0.28	0.32			
HD4H: 100.74	M	0.98	0.95	0.96	-1.36			
HD4H: 120.58	N, M+N					1.56	0.16	
HD5H: 29.9-33.09	M, N, M+N	0.13	0.70	0.41	-0.40	-0.48	-0.76	
HD5H: 63.31-67	N, M+N	1.00	0.97	0.99	1.41	1.96		
HD7H: 39- 45	M, N	0.07	0.45	0.25	-0.38	0.66		**X**
HD7H: 86.44	M	0.07	0.45	0.25	-1.42			**X**

[a]Location of QTL: xH:yy where x is the chromosome number and yy is the centimorgan position (cM).

[b]Allele frequency refers to the frequency of the "A" allele of the SNP marker with smallest p value.

[c]Allelic effect is effect of A allele relative to B allele.

[d] 'X' refers to QTL validated using the validation panel derived from crosses among selected parents from the University of Minnesota and North Dakota State University breeding programs. The cells with no 'X' indicate QTL that are not validated in our study.

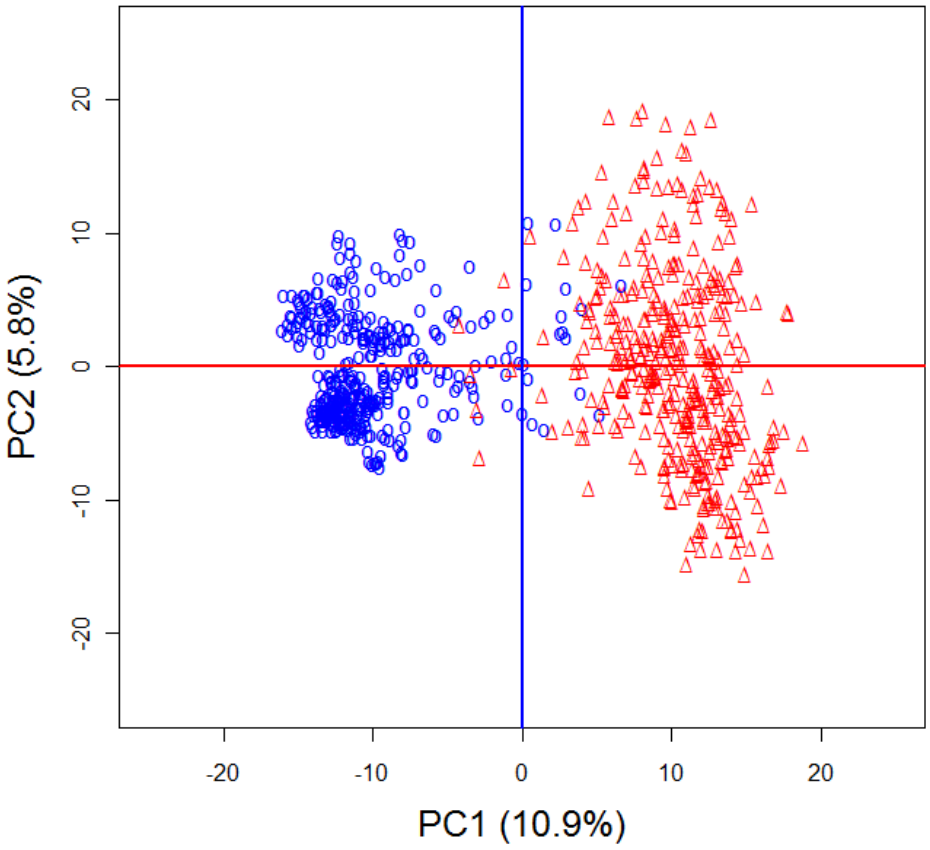

Fig.1 Plot of first (PC1) versus second (PC2) principal components for 384 lines from University of Minnesota (blue circles) and 384 lines from the North Dakota State University (red triangles) breeding programs using 2,110 SNP markers for each line.

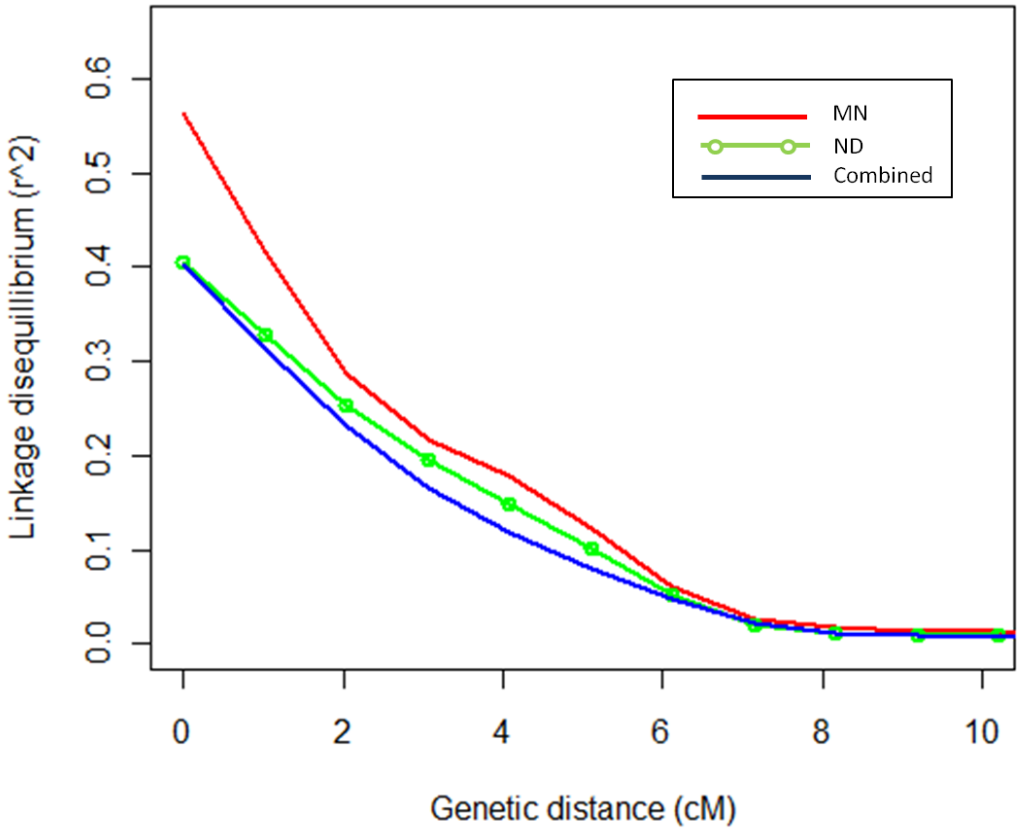

Fig.2 Decay of linkage disequilibrium in six-row spring barley germplasm estimated from Minnesota (red), North Dakota (green) and combined populations (blue). The smooth LD curves were fitted using LOWESS local regression package in R.

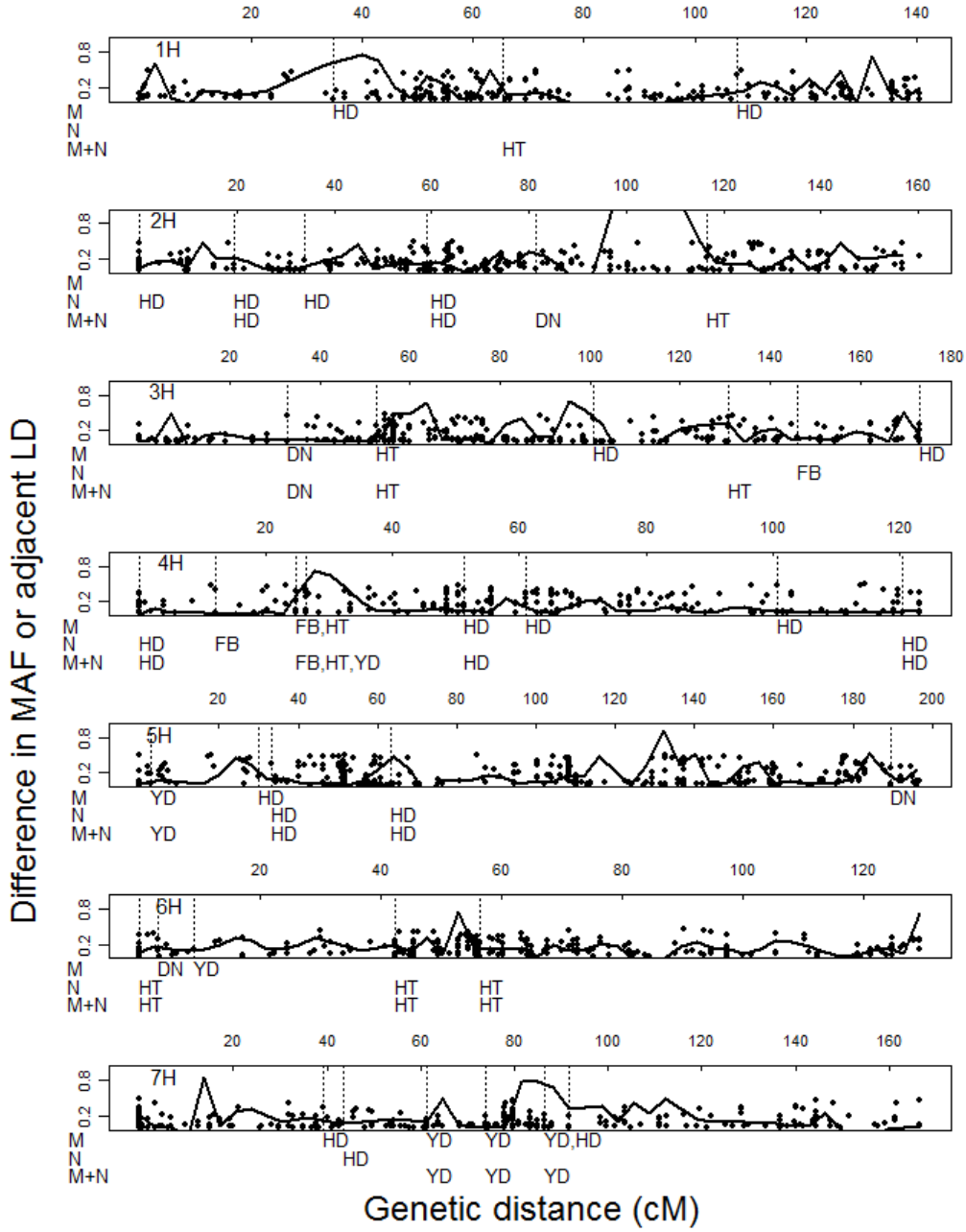

Fig.3 Distribution of absolute difference in minor allele frequency (MAF; black solid circles) and absolute difference in adjacent marker linkage disequilibrium (LD; black solid line) between the Minnesota and North Dakota breeding programs based on 2,110 single nucleotide polymorphisms. Vertical dotted lines indicate the position of a QTL which is labeled below as being detected in the Minnesota (M), North Dakota (N) or combined (M+N) mapping panels. Traits for QTL are labeled as Fusarium head blight severity (FH), deoxynivalenol concentration (DN), grain yield (YD), heading date (HD), and plant height (PH). Genetic distance (cM) is indicated above each chromosome panel.

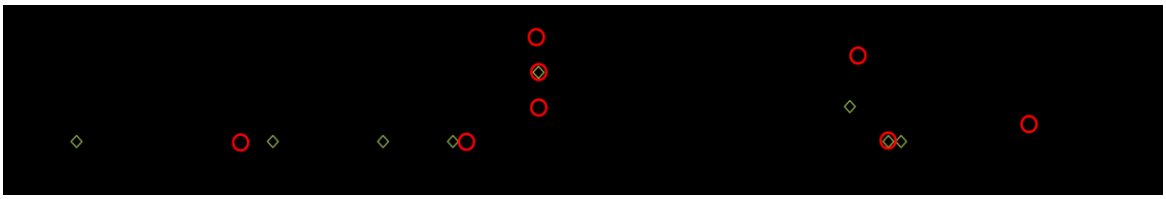

Fig.4 A summary of QTL identified from genome wide association mapping for FHB (square), DON (triangle), plant height (diamond), grain yield (solid circle) and heading date (open circle) in six-row spring barley germplasm (See supplementary figure S1 to S5 for detailed Manhattan plots). For genomic locations associated with multiple traits, symbols are stacked along the y axis for clarity. The symbols circled in red refer to QTL that were validated using the validation mapping panel.

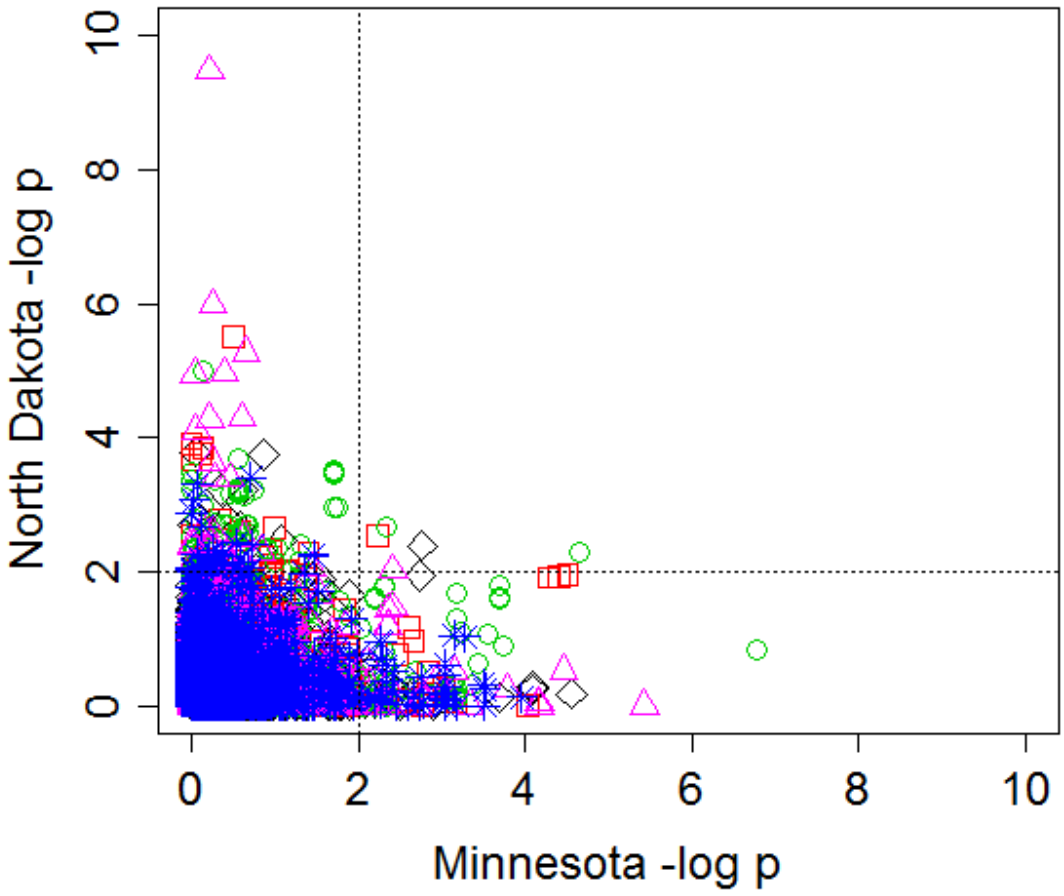

Fig.5 Relationship between the level of significance (–logP value) of a SNP marker mapped using the Minnesota and North Dakota mapping panels for all five traits (FHB = square, DON = diamond, heading date = circle, plant height = triangle and grain yield = star). Dotted lines represent a -logp threshold of two and show that very few SNPs were significant in both breeding programs.

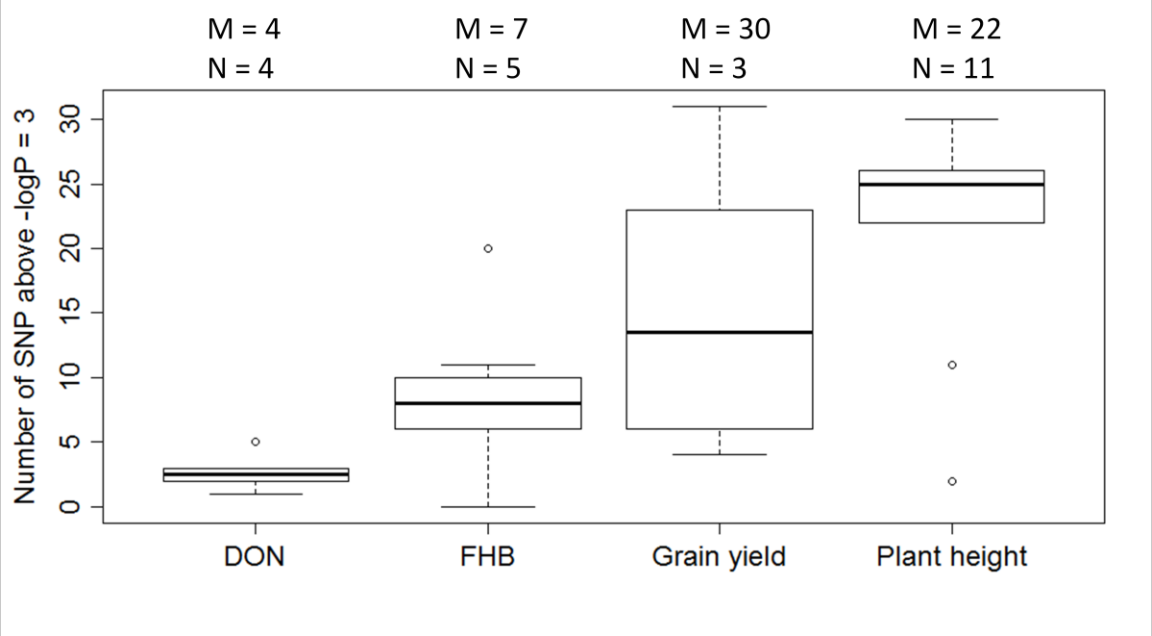

Fig.6 Number of SNP detected above -logp threshold of three identified in M+N:384 using random re-sampling analyses. Numbers on the top of each box plot indicate number of SNP above -logp threshold of three identified in M and N mapping panels with same sample size (N=384). For heading date, an average of 45 SNP (Min=11, Max=91) exceeded our -logp threshold of three for M+N:384 resampled data set. Thirty three and 53 SNP exceeded our -logpF threshold for M and N. The heading date information was not included in figure to keep the clarity of distribution for other traits.

Reference

Beavis, W.D., 1998. QTL analyses: power, precision, and accuracy, pp. 145–162 in Molecular Dissection of Complex Traits, edited by A.H. Paterson. CRC Press, Boca Raton,

Belknap, J. K., Mitchell, S. R., O'Toole, L. A., Helms, M. L., & Crabbe, J. C. (1996). Type I and type II error rates for quantitative trait loci (QTL) mapping studies using recombinant inbred mouse strains. Behavior genetics, 26(2), 149-160.

Bernardo R (2008) Molecular markers and selection for complex traits in plants: learning from the last 20 years. Crop Sci 48:1649-1664

Blake V, Kling J, Hayes P, Jannink J-L, Jillella S, Lee J, Matthews D, Chao S, Close T, Muehlbauer G, Smith K, Wise R, Dickerson J (2012) The Hordeum toolbox—The barley CAP genotype and phenotype resource. Plant Genome 5:81–91

Bradbury P, Parker T, Hamblin MT, Jannink JL (2011) Assessment of power and false discovery in genome-wide association studies using the BarleyCAP germplasm. Crop Sci 51:52-59

Bradbury P, Parker T, Hamblin MT, Jannink JL (2011) Assessment of power and false discovery rate in genome-wide association studies using the BarleyCAP germplasm. Crop Sci 51:52–59

Breseghello, F., & Sorrells, M. E. (2006). Association mapping of kernel size and milling quality in wheat (Triticum aestivum L.) cultivars. Genetics, 172(2), 1165-1177.

Bush WS, Moore JH (2012) Chapter 11: Genome-Wide Association Studies. PLoS Comput Biol 8(12): e1002822. doi:10.1371/journal.pcbi.1002822

Canci, P. C., Nduulu, L. M., Muehlbauer, G. J., Dill-Macky, R., Rasmusson, D. C., & Smith, K. P. (2004). Validation of quantitative trait loci for Fusarium head blight and kernel discoloration in barley. Molecular Breeding, 14(2), 91-104.

Carlson, C. S., Eberle, M. A., Kruglyak, L., & Nickerson, D. A. (2004). Mapping complex disease loci in whole-genome association studies. Nature, 429(6990), 446-452.

Carlson, C. S., Eberle, M. A., Kruglyak, L., & Nickerson, D. A. (2004). Mapping complex disease loci in whole-genome association studies. Nature, 429(6990), 446-452.

Close TJ, Bhat PR, Lonardi S, Wu Y, Rostoks N, Ramsay L, Druka A, Stein N et al (2009) Development and imple-mentation of high-throughput SNP genotyping in barley. BMC Genomics 10:582

Comadran J, Ramsay L, MacKenzie K, Hayes P, Close TJ, Muehlbauer G, Stein N, Waugh R.2011. Theor Appl Genet. 2011 Feb;122(3):523-31.

Comadran, J., Russell, J. R., Van Eeuwijk, F. A., Ceccarelli, S., Grando, S., Baum, M., ... & Thomas, W. T. B. (2008). Mapping adaptation of barley to droughted environments. Euphytica, 161(1-2), 35-45.

Condón[a], F., C. Gustus , D.C. Rasmusson, and K.P. Smith. 2008. Effect of Advanced Cycle Breeding on Genetic Diversity in Barley Breeding Germplasm. Crop Sci. 48:1027–1036.

Condon[b], F., D.C. Rasmusson1, E. Schiefelbein G. Velasquez, and K.P. Smith. 2008. Effect of Advanced Cycle Breeding on Genetic Gain and Phenotypic Diversity in Barley Breeding Germplasm. Crop Sci. 49: 1751-1761.

Dahleen, L. S., Agrama, H. A., Horsley, R. D., Steffenson, B. J., Schwarz, P. B., Mesfin, A., & Franckowiak, J. D. (2003). Identification of QTLs associated with Fusarium head blight resistance in Zhedar 2 barley. Theoretical and applied genetics, 108(1), 95-104.

De la Pena, R. C., Smith, K. P., Capettini, F., Muehlbauer, G. J., Gallo-Meagher, M., Dill-Macky, R., ... & Rasmusson, D. C. (1999). Quantitative trait loci associated with resistance to Fusarium head blight and kernel discoloration in barley. Theoretical and Applied Genetics, 99(3-4), 561-569.

Dorweiler, J., Stec, A., Kermicle, J., & Doebley, J. (1993). Teosinte glume architecture 1: a genetic locus controlling a key step in maize evolution. Science, 262(5131), 233-235.

Druka, A., Franckowiak, J., Lundqvist, U., Bonar, N., Alexander, J., Houston, K., ... & Waugh, R. (2011). Genetic dissection of barley morphology and development. Plant physiology, 155(2), 617-627.

Feuillet C, Langridge P, Waugh R. Cereal breeding takes a walk on the wild side. Trends in Genetics 2008;24:24-32.

Gebhardt, D.J., D.C. Rasmusson, and R.D. Wilcoxson. Cyclic breeding to incorporate kernel discoloration resistance into malting barley. Crop Sci. 1992. 32:352–356.

Goblirsch, C.A., R.D. Horsley, and P.B. Schwarz. A strategy to breed low-protein barley with acceptable kernel color and diastatic power. Crop Sci. 1996. 36:41–44

Hajjar, R., & Hodgkin, T. (2007). The use of wild relatives in crop improvement: a survey of developments over the last 20 years. Euphytica, 156(1-2), 1-13.

Hayes, B. J., & Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. Genetics, 157(4), 1819-1829.

Horsley, R. D., Schmierer, D., Maier, C., Kudrna, D., Urrea, C. A., Steffenson, B. J., ... & Kleinhofs, A. (2006). Identification of QTLs associated with Fusarium head blight resistance in barley accession CIho 4196. Crop science, 46(1), 145-156.

Horsley, R. D., Schwarz, P. B., & Hammond, J. J. (1995). Genetic diversity in malt quality of North American six-rowed spring barley. Crop science, 35(1), 113-118.

Huang Y, Millett BP, Beaubien KA, Dahl SK, Steffenson BJ, Smith KP, Muehlbauer GJ. (2012). Haplotype diversity and population structure in cultivated and wild barley evaluated for Fusarium head blight responses. Theor. Appl. Genet. 126 (3): 619-36.

Ioannidis, J.P., Thomas, G. & Daly, M.J. Validating, augmenting and refining genome-wide association signals. Nat. Rev. Genet. 10, 318–329 (2009).

Jannink, J. L., & Walsh, B. (2002). Association mapping in plant populations. Quantitative genetics, genomics and plant breeding, 59-68.

Kang et al. (2008) Efficient control of population structure in model organism association mapping. Genetics. 2008 Mar;178(3):1709-23.

King, E. G., C. M. Merkes, C. L. McNeil, S. R. Hoofer, S. Senet al.,2012 Genetic dissection of a model complex trait using the Drosophila Synthetic Population Resource. Genome Res. 22

Kover, P. X., Valdar, W., Trakalo, J., Scarcelli, N., Ehrenreich, I. M., Purugganan, M. D., ... & Mott, R. (2009). A multiparent advanced generation inter-cross to fine-map quantitative traits in Arabidopsis thaliana. PLoS genetics, 5(7), e1000551.

Lander E.S., and Kruglyak L. (1995) Genetic dissection of complex traits: Guidelines for interpreting and reporting linkage results. Nat. Genet. 11:241–247.

Lin X.H., Zhang D.P., Xie Y.F., Gao H.P., Zhang, Q., 1996. Identifying and mapping a new gene for bacterial blight resistance in rice based on RFLP markers. Phytopathology 86: 1156-1159.

Long AD, Langley CH. 1999. The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits. Genome Res. 1999 Aug;9(8):720-31.

Lundqvist, U. & Lundqvist, A. Induced intermedium mutants in barley: origin, morphology and inheritance. Hereditas 108, 13–26 (1988).

Massman J, Cooper B, Horsley R, Neate S, Dill-Macky R, Chao S, Dong Y, Schwarz P, Muehlbauer GJ, Smith KP (2010) Genome-wide association mapping of Fusarium head blight resistance in contemporary barley breeding germplasm. Mol Breed doi: 10.1007/s11032-010-9442-0

Mesfin, A., Smith, K. P., Dill-Macky, R., Evans, C. K., Waugh, R., Gustus, C. D., & Muehlbauer, G. J. (2003). Quantitative trait loci for Fusarium head blight resistance in barley detected in a two-rowed by six-rowed population. Crop science, 43(1), 307-318.

Nduulu, L. M., Mesfin, A., Muehlbauer, G. J., & Smith, K. P. (2007). Analysis of the chromosome 2 (2H) region of barley associated with the correlated traits Fusarium head blight resistance and heading date. Theoretical and Applied Genetics, 115(4), 561-570.

Peel, M. D., & Rasmusson, D. C. (2000). Improvement strategy for mature plant breeding programs. Crop science, 40(5), 1241-1246.

Pillen-K, Zacharias-A, Leon-J, "Advanced backcross QTL analysis in barley (Hordeum vulgare L.)", Theor Appl Genet, 2003, vol. 107, pp. 340-352

Pritchard JK, Cox NJ. 2002. The allelic architecture of human disease genes: Common disease–common variant…or not? *Hum Mol Genet* 11*:* 2417–2423.

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. Genetics, 155(2), 945-959.

Pumphrey, M. O., Bernardo, R., & Anderson, J. A. (2007). Validating the QTL for Fusarium head blight resistance in near-isogenic wheat lines developed from breeding populations. Crop Science, 47(1), 200-206.

Rafalski, A. (2002). Applications of single nucleotide polymorphisms in crop genetics. Current opinion in plant biology, 5(2), 94-100.

Ramsay, L., Comadran, J., Druka, A., Marshall, D. F., Thomas, W. T., Macaulay, M., ... & Waugh, R. (2011). INTERMEDIUM-C, a modifier of lateral spikelet fertility in barley, is an ortholog of the maize domestication gene TEOSINTE BRANCHED 1. Nature genetics, 43(2), 169-172.

Rasmusson DC, Phillips RL. Plant breeding progress and genetic diversity from de novo variation and elevated epistasis. Crop Sci. 1997;37(2):303–310.

Rostoks N, Mudie S, Cardle L, Russell J, Ramsay L, Booth A, Svens-son JT, Wanamaker SI, Walia H, Rodriguez EM, Hedley PE, Liu H, Morris J, Close TJ, Marshall DF, Waugh R (2005) Genome-wide SNP discovery and linkage analysis in barley based on genes responsive to abiotic stress. Mol Gen Genet 274:515–527

Scho¨n CC, Utz HF, Groh S, Truberg B, Openshaw S, Melchinger AE (2004) Quantitative trait locus mapping based on resampling in a vast maize testcross experiment and its relevance to quantitative genetics for complex traits. Genetics 167:485–498

Smith, K. P., Budde, A., Dill-Macky, R., Rasmusson, D. C., Schiefelbein, E., Steffenson, B., Wiersma, J. J., Wiersma, J. V., & Zhang, B. (2013). Registration of 'Quest'spring malting barley with improved resistance to Fusarium head blight. Journal of Plant Registrations, 7(2), 125-129.

Tester, M., & Langridge, P. (2010). Breeding technologies to increase crop production in a changing world. Science, 327(5967), 818-822.

van Berloo, R., Aalbers, H., Werkman, A., & Niks, R. E. (2001). Resistance QTL confirmed through development of QTL-NILs for barley leaf rust resistance. Molecular Breeding, 8(3), 187-195.

Wang H, Smith KP, Combs E, Blake T, Horsley RD et al (2011) Effect of population size and unbalanced data sets on QTL detection using genome-wide association mapping in barley breeding germplasm. Theor Appl Genet

Weir, B. S., & Cockerham, C. C. (1984). Estimating F-statistics for the analysis of population structure. evolution, 1358-1370.

Wu¨rschum T, Maurer HP, Kraft T, Janssen G, Nilsson C et al (2011a) Genome-wide association mapping of agronomic traits in sugar beet. Theor Appl Genet 123:1121–1131

Wu¨rschum T, Maurer HP, Schulz B, Mo¨hring J, Reif JC (2011b) Genome-wide association mapping reveals epistasis and genetic interaction networks in sugar beet. Theor Appl Genet 123:109–118

Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Heath AC, Martin NG, Montgomery GW, et al. 2010. Common SNPs explain a large proportion of the heritability for human height. *Nat Genet* 42*:* 565–569.

Young, N. D. (1996). QTL mapping and quantitative disease resistance in plants. Annual review of phytopathology, 34(1), 479-501.

Young, N. D. (1999). A cautiously optimistic vision for marker-assisted breeding. Molecular breeding, 5(6), 505-510.

Yu, J., Pressoir, G., Briggs, W. H., Bi, I. V., Yamasaki, M., Doebley, J. F., ... & Buckler, E. S. (2005). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nature genetics, 38(2), 203-208.

Zhou, H., & Steffenson, B. J. (2013). Genome-wide association mapping reveals genetic architecture of durable spot blotch resistance in US barley breeding germplasm. In Advance in Barley Sciences (pp. 257-267). Springer Netherlands.

Zhu, C et al. (2010) Status and Prospects of Association Mapping in Plants. The Plant Genome 1, 5-20.

Zhu, H., Gilchrist, L., Hayes, P., Kleinhofs, A., Kudrna, D., Liu, Z., Prom, L., Steffenson, B., Toojinda, T & Vivar, H. (1999). Does function follow form? Principal QTLs for Fusarium head blight (FHB) resistance are coincident with QTLs for inflorescence traits and plant height in a doubled-haploid population of barley. Theoretical and Applied Genetics, 99(7-8), 1221-1232.

Chapter 2

Impact of training population composition on progeny prediction accuracy: genomic prediction using breeding lines from two North American barley improvement programs

Although several simulation and cross validation studies have indicated that genomic selection (GS) could have a substantial impact on breeding, the overall effectiveness of GS to improve breeding efficiency will be determined to a great degree by the accuracy of predictions of progeny performance. In addition, empirical studies using progeny derived from a training population are needed to design strategies for GS in breeding. To study the effect of training population composition on progeny prediction accuracy, we designed a GS study using six-row spring malting barley germplasm from two closely related breeding programs in the Upper Midwest. For model training, we designed three different panels using 384 breeding lines from the University of Minnesota (M), 384 lines from North Dakota State University (N), and 768 lines from both programs combined (M+N). Our prediction candidates consisted of 300 progeny lines derived from three different cross types from two breeding programs: MxM, NxN and MxN. Each cross type consisted of 100 F_4 derived progeny lines. All lines in the training population were genotyped with 3,072 SNP markers. Prediction candidates were genotyped at F_3 generation with 384 selected SNP markers and phenotyped for grain yield, plant height, heading date, DON accumulation, FHB severity and grain protein. Our results indicate that (1) the prediction accuracy of progeny for six traits were moderate to moderately high indicating that GS could be successfully implemented for agronomic, disease and quality traits with a range of heritabilities, (2) prediction accuracy was higher when M

was used to predict MxM or N to predict NxN, (3) M+N predicted MxN better than either M or N alone, (4) combining lines from breeding programs to increase training population size did not increase our prediction accuracy. We conclude that GS models hold excellent potential over phenotypic selection to improve breeding efficiency of yield and FHB disease resistance in barley – two traits that show complex inheritance and low to moderate heritability.

Introduction

Marker-assisted selection (MAS) and marker assisted recurrent selection (MARS) have been widely employed in plant breeding (Eathington et al. 2007); however, selection based on genome-wide markers is expected to be more efficient to improve traits with complex inheritance (Bernado 2008; Heffner et al. 2010; Lorenz et al. 2011). Genomic selection (GS) is a method to statistically predict breeding value of individuals for such complex traits by simultaneously fitting genome-wide markers on phenotype data (Meuwissen et al. 2001; Whittaker et al. 2000). Implementing GS in breeding has become increasingly more feasible due to rapidly evolving DNA marker and sequencing technology that has dramatically reduced costs and increased throughput. Simulation and cross-validation (CV) studies suggest that, for traits controlled by many small-effect genes, higher response to selection could be achieved by selecting lines based on the sum of the allelic effect of many markers spanning the whole genome compared to traditional MAS based on a few significant selected markers (Lorenzana and Bernado 2009; Heffner et al. 2011). Because GS identifies superior parents based on breeding values estimated from markers rather than phenotyping, GS could greatly reduce breeding cycle time and increase the rate of gain from selection (Bernado and Yu 2007). The ultimate success of GS in crop improvement will be determined in large part by the accuracy of predictions based on genome-wide markers.

Most studies aimed at assessing prediction accuracy of GS in plants have used either simulated data (Bernardo and Yu 2007; Zhong et al. 2009) or cross-validation approaches (De los Campos et al. 2009; Crossa et al. 2010; Asoro 2011; Heffner et al.

2011). By simulating two cycles of selection, Bernardo and Yu (2007) reported that response to GS could be 18% to 43% larger than the response to marker-assisted recurrent selection. Using a multi-environment cross-validation study in wheat and maize, Crossa et al. (2010) indicated that GS models that included marker data or marker and pedigree data performed better than models that included pedigree data alone, and that there was no significant difference between GS models that included marker data only or marker and pedigree data. Average prediction accuracy of GS was 28% greater than MAS in a multifamily cross-validation study using 374 winter wheat breeding lines (Heffner et al. 2011). This study emphasized the potential of GS to improve genetic gain per unit time and cost, and pointed out the need for further research to optimize GS in dynamic breeding programs. An empirical selection experiment using a maize bi-parental cross indicated that, for grain yield and stover quality traits, GS resulted in 14 to 50% larger gains than MARS; but, GS gains from cycle 1 to cycle 3 after an initial cycle of phenotypic selection were small and inconsistent (Massman et al. 2013). Additional empirical selection experiments using breeding data sets are needed to fully assess the potential of GS as a tool for breeders.

While cross-validation and simulation studies have been useful to investigate factors that influence genomic prediction accuracy, empirical studies predicting progeny performance will be more relevant to breeding programs seeking to implement GS. Validation studies using progeny derived from a training population have two major advantages over cross-validation studies. First, studying prediction accuracy on progeny populations will address the effect of recombination between markers and QTL occurring

during the breeding process. We expect prediction accuracy estimates from cross-validation studies to be inflated compared to estimates of prediction accuracy on progenies that have undergone recombination (Technow et al. 2013). Second, progeny performance will be tested in new environments as would be done in an actual breeding program; whereas, in cross-validation studies both training and validation populations share common test environments. This results in common genotype-by-environment effects that will bias the correlation between estimated and observed breeding values upwards as opposed to having independent training and validation panels tested in different environments (Lorenz et al. 2011). Therefore, empirical studies using progeny derived from a training panel are needed to better inform the design of strategies for GS in breeding.

The composition and size of the training panel can affect GS prediction accuracy. Previous studies in cattle and crops have reported that increasing the training panel size by combining subpopulations resulted in higher prediction accuracies than training panels with only one subpopulation (Hayes et al. 2009; Toosi etal 2010; Asoro et al. 2011; Technow et al. 2013). However, in barley, breeding lines from a single breeding program were better predictors of lines from that same program and pooling lines from other breeding programs to increase training panel size did not increase prediction accuracy (Lorenz et al. 2012). Combining data from multiple breeding populations to improve prediction accuracy may be valuable to breeders working with traits that are expensive to measure or when training population sizes are small. However, this strategy may be suboptimal or even detrimental if marker effects are very different between

subpopulations or predominant marker–QTL linkage phases change across subpopulations (Hayes et al. 2010; Lorenz et al. 2012).

The ability to predict progeny performance of crosses made among unrelated breeding programs, or universal prediction equations to predict line performance in multiple breeding populations can facilitate large collaborative breeding efforts using GS approaches. Breeders could make crosses among elite breeding lines from other programs to increase genetic variance in their breeding pool and improve selection efficiency without the issues of linkage drag associated with using exotic germplasm. Previously, making wide crosses even with elite breeding material required up to three breeding cycles to achieve modest gains (Peel and Rasmusson 2000). Such germplasm exchange, if coupled with GS could reduce breeding cycle time and accelerate the introgression of novel alleles into breeding populations. More research is needed to assess the potential of GS to improve breeding efficiency of using elite germplasm.

In this study, we investigate the accuracy of genomic predictions using progeny derived from crosses among parents from within and between two barley breeding programs from the Upper Midwest. We build on the discoveries made by Lorenz et al. (2012) that demonstrated the importance of training population composition on prediction accuracy for Fusarium head blight (FHB) resistance using a cross-validation study. Our objectives were to: (1) determine GS prediction accuracy on progenies and contrast them to previous cross-validation results; (2) investigate the effect of training population composition, sample size and model on prediction accuracy of FHB and other traits representing a range in heritability.

Materials and methods

We investigated accuracy of genomic predictions using a population comprised of 300 progeny lines derived from crosses among 14 parents from the University of Minnesota (MN) and North Dakota State University (ND) barley breeding programs. Eleven of the 14 parents are from a collection of 768 elite breeding lines from the MN and ND breeding programs, hereafter referred to as the parent population. Phenotypic data for the parent population were collected in trials conducted from 2006 – 2009 in Minnesota and North Dakota as part of the Barley Coordinated Agricultural Project (Barley CAP; www. barleycap.org). The data for the 300 progeny were collected in trials conducted in 2011 and 2012 at Minnesota and North Dakota. We assessed prediction accuracy for six traits using two prediction models.

Parent population

The parent population consisted of 384 lines from the MN and ND barley breeding programs. Over the four year period from 2006 – 2009, sets of 96 lines from each breeding program were evaluated and referred to as CAP1– 4. The study by Lorenz et al. (2012) utilized CAP1, 2, and 3 sets, while this study used all four CAP sets. F4 or F5 lines were derived through single-seed descent and evaluated in a randomized complete block design with one or two replications carried out at multiple test locations in Minnesota and North Dakota from 2006 to 2009. Repeated check cultivars were Baronesse, Harrington, Lacey, Legacy, MNBrite, Robust, Stander, Stellar, and Tradition in the Minnesota trials, and Baronesse, Harrington, and Robust in the North Dakota trials. For model training, we designed three different training panels using 384 breeding lines

from the University of Minnesota (M), 384 lines from North Dakota State University (N), and 768 lines from both programs combined (M+N).

Phenotype data: Phenotypic data for heading date, plant height, grain yield, FHB severity, DON concentration and grain protein were collected from the field trials (Table S1&S2). Heading date was measured as the number of days after planting in which at least 50% of the spikes in a plot were emerged at least half way from the boot. Plant height was measured as the length of the plant (cm) from the soil surface to the tip of the spike, excluding awns, and was recorded as the mean of two measurements per plot. Plots were harvested with a Wintersteiger Master Elite combine plot harvester (Wintesteiger, Ried, Germany) and grain yield was recorded as kilograms per hectare at 14 percent moisture content. Grain protein concentration was determined with a Diode Array 7250 near infrared reflectance (NIR) grain analyzer (Perten Instruments, Sweden). For disease traits, FHB and DON, entries were planted in single row plots, inoculated with *Fusarium graminearum* as either macroconidia or grain spawn, and mist irrigated following inoculum application. Disease severity was measured as the percent infected kernels based on visual ratings of 10 arbitrarily selected spikes and mature grain was harvested from plots, ground, and analyzed by gas chromatography and mass spectrometry as described by Jones and Mirocha (1999). The details of the experiment setup and disease evaluation protocols for FHB severity and DON accumulation are described by Massman et al. (2011). All phenotypic and genotypic (see below) data referenced in this manuscript could be downloaded from The T3 Barley database http://triticeaetoolbox.org/barley/) developed as a part of Triticeae Coordinated Agricultural Project (Blake et al. 2012).

Genotype data: The parent population was genotyped with 3,072 SNP markers as a part of Barley CAP project using two Illumina oligo pool assays (OPA) containing 1,536 SNPs each and referred to as BOPA1 and BOPA2 (Close et al. 2009). The details of genotyping procedure are described in Massman et al. (2011); briefly, DNA was extracted from leaf tissue harvested from 3 week old plants at F4 generation using the protocol of Slotta et al. (2008), and the Illumina BeadStation was used to genotype each line utilizing the Golden Gate assay, as described in Fan et al. (2006).

Statistical analysis: We checked all traits for obvious outliers based on bivariate plots of raw data. The residuals were checked for normality (via different normality tests using histogram, box plot, stem-and-leaf plots and QQ plot) and homoskedasticity (via residual plots) using PROC UNIVARIATE (SAS®, version 9.3.1). Observations with a studentized residual absolute value greater than three were removed from our analysis as outliers. We used PROC MEANS to calculate trait means.

Because our phenotype data were collected from many trials that spanned over four years and testing locations for two breeding programs, we employed a two-step analysis to calculate best linear unbiased estimates (BLUE) of individual lines. First, common repeated check varieties from all trials were used to estimate respective trial effects for all traits using the model $\mathbf{Y} = \mathbf{X\alpha} + \mathbf{C\beta} + \mathbf{e}$; where $\mathbf{Y}$ is the vector of check phenotypes, $\mathbf{X}$ is an incidence matrix relating trial effects to $\mathbf{Y}$; $\boldsymbol{\alpha}$ is a vector of trial fixed effects to be estimated; $\mathbf{C}$ is an incidence matrix relating check lines to $\mathbf{Y}$; $\boldsymbol{\beta}$ is a vector of check fixed effects; and $\mathbf{e}$ is a vector of random residuals where $e \sim \mathbf{N}(0, \sigma^2_e \mathbf{I})$. Second, each trait in each trial was adjusted for respective trial effects, and best linear unbiased

estimates (BLUE) of individual lines were calculated using the model $\mathbf{Y_{adj}} = \mathbf{L\gamma} + \mathbf{e}$; where $\mathbf{Y_{adj}}$ is the vector of trait adjusted for corresponding trial effects. **L** is an incidence matrix relating line effects to $\mathbf{Y_{adj}}$; $\boldsymbol{\gamma}$ is a vector of line effects to be estimated; and **e** is a vector of random residuals where $\mathbf{e} \sim \mathbf{N}(0, \sigma^2_e \mathbf{I})$. We trained GS models using the BLUE of individual lines. We used PROC MIXED to estimate variances (SAS®, version 9.3.1) from Y_{adj}. Broad sense heritability on an entry means basis was calculated from these variances using the model $H = \sigma^2_L / [\sigma^2_L + \sigma^2_e/r]$; where σ^2_L is the line variance, σ^2_e is the error variance, r is the number of trials a line was tested.

Progeny population

The progeny population consisted of sets of 100 progeny lines from each of three different cross types: (1) crosses among parents within the MN program (MxM), (2) crosses among parents within the ND program (NxN), and (3) crosses using parents from both breeding (MxN). Fourteen advanced breeding lines from the MN and ND programs served as parents for the cross types. For each cross type, ten crosses were generated and twenty-four lines per cross were advanced by single seed decent to the F_3 generation (i.e., 240 lines/cross type). Progeny lines were genotyped at F3 generation (see below). F3:4 head rows were grown in a winter nursery in New Zealand to produce F3:5 seed for replicated field trials. One hundred lines per cross type with at least 10 random lines per cross were selected from each cross type to comprise the progeny population of 300 lines.

Phenotype data: The phenotypic data for the 300 progeny lines were collected from field trials in 2011 and 2012 using an augmented block design (Federer, 1956) with six blocks. In each block, we tested fifty progeny lines along with three check varieties that are

replicated twice. Agronomic and disease related trait data were collected in separate trials. Agronomic traits were evaluated at St Paul, MN, Crookston, MN, and Nesson Valley, ND in 2011 and in St Paul, MN, Crookston, MN, Nesson Valley, ND and Fargo, ND in 2012. The plot sizes and seeding rates were 1.9 m^2 and 330 plants/m^2 for trials at MN, and 3.7 m^2 and 247 plants/m^2 for trials at ND. For disease trials, we followed the same augmented block design except that the trial was replicated twice at each location. Disease traits were evaluated in St Paul, MN, Crookston, MN, Osnabrock, ND and Langdon, ND in 2011 and in St Paul, MN, Crookston, MN, Osnabrock, ND and Langdon, ND in 2012. FHB and DON were evaluated in single row plots 1.5 m in length planted with 4 g of seed per row at MN and hill plots at ND with 12 seeds per hill. Three common checks used were Tradition, Quest, and Lacy for yield trials and Tradition, Quest, and ND20493 for disease trials. All phenotypes were measured as described for the parent population above.

Genotype data: The 384 SNPs for marker predictions were selected from barley OPA1 and OPA2 based on polymorphic information content among the 14 parents and even distribution across genome (Close et al. 2009). We developed an Illumina BeadExpress Custom Array with the 384 selected SNP markers (Table S4). We observed that 21 markers failed during progeny genotyping and hence they were removed. Similarly, markers were removed if they had minor allele frequencies less than 0.001 or excessive heterozygosity prior to analysis. None of the heterozygous markers had frequency above expected level (1/16). After quality control, our marker data set consisted for 340

markers. All missing markers were replaced with a heterozygous score for prediction analysis.

Statistical analysis: We removed outliers and checked for normality and homoskedasticity of the residuals as described above. Additionally, a trial was removed if it reduced overall pooled entry-mean heritability (see below). For example, two FHB trials at North Dakota performed poorly in 2012. Pooled entry-mean heritability estimates of FHB with and without these two trials indicated these trials introduced substantial noise to the overall data set and hence both were removed from any subsequent analysis. We used PROC MEANS to calculate trait means (SAS®, version 9.3.1). We used principal component scores of cultivars estimated using SNP markers to determine sub population grouping. For agronomic traits, a total of three to seven trials were used and for DON and FHB four to six trials were used (Table S3).

A two-stage analysis was used to calculate BLUE of progeny lines. First, common repeated check varieties from each block were used to estimate block effects for all traits and trials using the model $\mathbf{Y} = \mathbf{X\alpha} + \mathbf{C\beta} + \mathbf{e}$; where $\mathbf{Y}$ is the vector of check phenotypes, $\mathbf{X}$ is an incidence matrix relating block effects to $\mathbf{Y}$; $\boldsymbol{\alpha}$ is a vector of block fixed effects of to be estimated; $\mathbf{C}$ is a matrix of check line information; $\boldsymbol{\beta}$ is a vector of check effects; and e is a vector of residual variation where $\mathbf{e} \sim \mathbf{N}(0, \sigma^2_e \mathbf{I})$. Second, each trait in each block was adjusted for respective block effects, if block effects were significant ($p<0.05$). For agronomic traits, BLUE of individual lines were calculated using the model $\mathbf{Y_{adj}} = \mathbf{T\beta} + \mathbf{L\gamma} + \mathbf{e}$; where $\mathbf{Y_{adj}}$ is the vector of trait adjusted for corresponding block effects in a trial, $\mathbf{L}$ is an incidence matrix relating line effect to $\mathbf{Y_{adj}}$; $\boldsymbol{\gamma}$ is a vector of fixed effect of

breeding lines to be estimated; **T** is an incidence matrix relating trial effect to $\mathbf{Y_{adj}}$; **β** is a vector of fixed effect of trials and **e** is a vector of residual variation where $e \sim \mathbf{N}(0, \sigma^2_e \mathbf{I})$. We used PROC MIXED to estimate variances (SAS®, version 9.3.1) and broad sense heritability on an entry means basis was calculated from these variances using the model $H = \sigma^2_L / [\sigma^2_L + \sigma^2_e/n]$; where n is number of trials. For FHB and DON traits, BLUE of individual lines were calculated using the model $Y_{adj} = T\beta + L\gamma + R\theta + e$; where R is a matrix of rep information; θ is a vector of fixed effect of rep and L, γ, T, and β are as described above. The broad-sense heritability on an entry means basis was calculated from these variances using the model $H = \sigma^2_L / [\sigma^2_L + \sigma^2_e/rn]$; where σ^2_L is the line variance, σ^2_e is the error variance, r is number of reps and n is number of trials.

Genomic selection models and predictions

We used the random regression best linear unbiased prediction (RR-BLUP ; Meuwissen et al. 2001) and Gaussian kernel method to perform GS analysis using a mixed model approach implemented in R-package rrBLUP (Endelman 2011). The M+N panel was used to train genomic prediction models and the marker estimates derived from the models were applied to the progeny marker genotypes to calculate predictions for each trait. The progeny prediction accuracy, here after referred to as 'prediction accuracy' was defined as the correlation between the predicted progeny value based on markers and the BLUE of progeny performance from multiple disease and yield trials. To account for the error contained within the entry means of the progeny, relative prediction accuracy was estimated by dividing prediction accuracy by the square root of the validation set entry mean heritability. Relative prediction accuracy facilitates comparison among other

related studies and among traits with different genetic architectures and heritabilities. A nonparametric estimate of the standard error of correlation was calculated using the bootstrap procedure with 1000 bootstrap replicates as implemented by Lorenz et. al (2012) and Efron and Gong (1983).

Training panel composition study: To study the effect of training panel composition on prediction accuracy, we divided the training and validation panels based on breeding programs. The scheme for estimating effect of training panel composition was as follows: The M, N, and M+N training panels were used to train models and these models were then applied to lines in the MxM, NxN, MxN progeny cross types to generate predictions. The predictions from three cross types were correlated to the BLUE of corresponding lines and relative prediction accuracies were calculated as described above.

To account for differences in population size when combining the M and N lines in a single training panel, we generated a M+N panel (n=384) through a systematic random re-sampling analysis without replacement. We generated 10 random re-sampled M+N data set with 384 lines (here after M+N:384) such that the contribution of M and N breeding lines was 1:1. The re-sampled M+N:384 data set was used to develop genomic predictions as described above.

Training panel sample size study: We studied the relative prediction accuracy as function of varying training panel sample size through random sampling without replacement. We applied a training panel size of 50, 150, 250 and 350 for M and N; and 50, 150, 250, 350, 550 and 750 for M+N training panels such that the contribution of M and N breeding lines was 1:1. For each training population size, 1000 random training

panels were generated, and each was used to generate RR-BLUP model to predict progeny line performance. M was used to predict 100 MxM progeny, N was used to predict 100 NxN progeny, and M+N was used to predict all 300 progeny lines that consisted of MxM, MxN and NxN progeny.

Comparison of progeny with cross-validation estimates of prediction accuracy: We performed two cross-validation analyses and a progeny prediction analysis. Cross-validation analyses were conducted on the parent population (M+N) and the progeny population after dividing them into training panels (n= 280 lines) and validation panels (n= 20 lines). The relative prediction accuracy from both cross-validation studies, here after referred to as 'cross-validation accuracy', was estimated through a 10 round resampling procedure without replacement. To maintain the same population size for the progeny prediction accuracy, a re-sampling procedure was used to generate a training set with 280 lines from M+N training panel and progeny data set with 20 lines from the progeny population.

Results

Variances and heritabilities for plant height, heading date, grain yield, FHB, DON and grain protein from parent and progeny populations are shown in Table 1 and Table 2 respectively. In the parent population, the trait heritabilities were similar in both M and N panels, except for heading date, where MN breeding lines had lower genetic variance and heritability than ND breeding lines. Similarly, MxM progeny also had much lower genetic variance and heritability for heading date than MxN or NxN progeny (Table 2). The genetic diversity of the progeny panels was further characterized using principal

component analysis using SNP marker data (Fig. 1). The first principal component explained most of the variation (95%) and separated the lines into the three different cross types with the MxN progeny falling in between the MxM and NxN progeny.

Relative prediction accuracy of progeny

Relative prediction accuracy ranged from 0.36 to 0.58 among the six traits included in our study (Fig. 2). FHB had the highest relative prediction accuracy, followed by heading date, DON, plant height, grain protein, and grain yield. On the other hand, heading date had the highest prediction accuracy, followed by FHB, plant height, DON, grain protein, and grain yield. As expected the difference between relative prediction accuracy and prediction accuracy was minimal when heritability was high.

Effect of training panel composition on genomic selection accuracy

To study the effect of training panel composition, we compared the relative prediction accuracy of four training panels to predict three progeny cross types. In 8 out of 10 cases, relative prediction accuracies were higher when M was used to predict MxM progeny or N to predict NxN progeny (Fig. 3). Surprisingly in the case of grain yield, N predicted MxM better than M. For four of the five traits, the combined panel (M+N) predicted progeny from crosses between programs (MxN) with equal or greater accuracy than either of the single program panels (M or N). When the M+N panel was used to predict the MxM or NxN progeny, it had greater accuracy compared to using either of the single program panels (M or N) in five of ten instances. In addition to changing the composition of the training panel, the combined panel (M+N) also had twice the number of individuals. Despite the difference in size, the relative prediction accuracy of the

combined data set with 384 lines (M+N:384) was very similar to M+N having 768 lines (Fig. 3).

Effect of training panel sample size on genomic selection accuracy

The negligible difference between M+N:384 and M+N:768 suggested that we reached a plateau for accuracy beyond a panel size of 384 (Fig. 3). To test this, we studied relative prediction accuracy as a function of training panel size using M to predict 100 MxM progeny, N to predict, 100 NxN progeny, and M+N training panel to predict all 300 progeny lines (Fig. 4). Our results indicated that prediction accuracy increased with training panel size; however, the gains in accuracy were small for large sample sizes. When M was used to predict MxM progeny or N to predict NxN progeny, no increase in accuracy was observed beyond a training panel sample size of 250. However, for combined panel (M+N) predicting all 300 progeny lines, accuracy continued to increase and reached a plateau only by a sample size of 550.

Comparison of progeny prediction and cross-validation accuracy

In general, the cross-validation accuracies were much greater than accuracies based on the parent-training panel predicting the progeny (Fig. 5). On an average, the cross-validation accuracies based on both parent and progeny sets were 48% to 58% higher than the realized progeny prediction accuracies. Interestingly, cross-validation accuracies observed in the progeny population was higher than cross-validation in the parent population for all of the traits except DON.

The GAUSS model often did much better than RR-BLUP for the cross-validation, but both models performed similarly for progeny prediction (Fig. 5). When we

considered the 12 training panel by progeny panel combinations for the five traits (n= 60), the RR-BLUP model and Gaussian kernel method performed similarly ($p = 0.03$ by two-sided pairwise t-test). A regression analysis also indicated that both RR-BLUP on GAUSS model accuracies were strongly correlated ($r^2 = 0.86$) and intercept was not significant ($p = 0.50$; Fig. 6). Considering each breeding program separately, the prediction accuracies between the two models were not different from each other based on a t-test ($p > 0.5$).

Discussion

The overall effectiveness of GS to improve breeding efficiency will be determined to a great degree by the accuracy of predictions of progeny performance. Previous studies using cross-validation indicate that prediction accuracies are sufficient to improve gain from selection given that the breeding cycle time will be shortened substantially as well (Bernado 2008; Heffner et al. 2010; Lorenz et al. 2012). In our study, the prediction accuracy of progeny for six traits were moderate to moderately high indicating that GS could be successfully implemented for agronomic, disease and quality traits with a range of heritabilities. To account for the error contained in the progeny phenotypic values, we calculated the relative prediction accuracy by dividing prediction accuracy by square root of heritability (h) within the validation set (Legarra et al. 2008). The relative prediction accuracy was higher than the prediction accuracy, suggesting that GS models holds excellent potential over phenotypic selection to improve breeding efficiency of yield and FHB that show complex inheritance and low to moderate heritabilities.

Assessment of genomic selection prediction accuracy using cross-validation should be interpreted cautiously

The accuracy of predictions for progeny performance is more relevant for a breeder planning to implement GS because in practice selection candidates are developed from crosses among superior lines in the previous generation and phenotyped in new environments. We found that the accuracy of predictions for progenies was lower than prediction accuracies based on cross-validation for all five traits studied. Similarly, the relative prediction accuracies for FHB and DON based on cross-validation in our previous study were also higher (r_p = 0.67 and 0.72; Lorenz et al. 2012) indicating that studies using this approach should cautiously speculate on the effectiveness of GS to predict genetic gains in subsequent breeding cycles. Hofheinz et al. (2012) similarly found that cross-validation prediction accuracy was not a good indicator of prediction accuracy in subsequent breeding cycles in sugar beets.

Crossing parents involves extra rounds of meiosis that can disrupt linkage blocks and change the linkage phase between markers and QTL in the resulting progeny. Cross validation studies, on the other hand, randomly split data sets into prediction and validation panels. Although these panels are assumed to be independent, in practice, lines in both panels may be related and are evaluated in the same environments. Furthermore, cross-validation results might be a biased indicator of actual progeny performance due to model over-fitting, a phenomenon where statistical models tend to remember the trends in data rather than learning to generalize underlying relationship between predictor and response variable (Hofheinz et al. 2012). Therefore, when cross-validation studies are

used to predict selection response, care must be taken to address the limitations of this methodology.

Optimizing genomic selection methods to improve accuracy

Because we found that the prediction accuracy of traits were moderate to moderately high, we were interested to explore the potentials of improving prediction accuracies and optimize our genomic selection schemes. We explored the effect of training panel composition, training panel size, and modeling methods on progeny prediction accuracy. Previous cross-validation studies indicated that prediction accuracies were better when training and prediction panels were closely related (Asoro et al. 2012; Lorenz et al. 2012). Lorenz et al. (2012) speculated that across breeding program predictions might be lower than within program predictions if segregating QTL are different, marker-QTL linkage phases are different or non-additive gene action exist within these breeding programs. Association mapping using the M and N training panels indicated that often different QTL were segregating for the traits included in our prediction models (See chapter 1). The same study also indicated that the two breeding programs differed in adjacent marker LD and allele frequencies. Differences in allele frequency between the two breeding populations could be the result of differential selection pressure or drift. In the extreme case where a marker is fixed in one population, that marker is no longer useful to predict allelic effects. Similarly, marker effect estimates of low frequency alleles in a population could be imprecise and this could potentially compromise the prediction accuracy if applied to a different population. Furthermore, we observed differences in marker effect estimates between the two breeding programs (Fig. S2). This

could be caused by differential recombination between the marker and QTL in these different populations. If so, using a higher density marker panel may increase the likelihood that a marker linked to a QTL is in the same phase in both populations. All these factors could affect marker effect estimates and lead to lower prediction accuracies when using one breeding population as a training panel to predict another breeding population. Therefore, breeders should design training panels that are closely related to their selection candidates.

Presence of non-additive gene action in advanced cycle MN breeding germplasm (Rassmuson and Philips 1997) and barley in general (Xu and Jia 2007; Li et al. 2008) have been previously suggested. To study the importance of non-additive gene action, we tested and compared accuracies of RR-BLUP and GAUSS models. If non-additive gene action is present, we hypothesized that a non-additive kernel based GAUSS model will perform better than a simple additive gene action based RR-BLUP model. It should be noted that, although kernel models will not capture all epistatic interactions directly, they can capture some non-additive gene action (Gianola and van Kaam 2008; Endelman 2011). Overall, both models produced similar prediction accuracies (Fig. 6) and no difference in progeny predictions accuracies among models were observed. Our findings largely indicate that most of the genetic variation in six row malt barley is additive in nature. It is possible that larger populations than those used in our study would be necessary to detect non-additive gene action, so this should be interpreted cautiously. Interestingly, the GAUSS model was more accurate than RR-BLUP in the

cross-validation analyses suggesting that non-additive models might be picking up some interaction effects present in the same dataset.

There were diminishing returns in accuracy for training panel sizes beyond 250 within a single breeding program similar to the findings of Lorenz et al. (2012). In contrast, the prediction accuracy continued to increase with sample size in the combined panel (M+N) highlighting the importance of maintaining larger training panel sizes when using a more diverse training panel. Improved prediction accuracy from larger training population is largely a function of effective population size (Ne) (Meuwissen et al. 2013; Daetwyler et al. 2010). If *Ne* is small, as in the case of most breeding populations, the number of independently segregating chromosome segments in a population will be small (Goddard 2009; Daetwyler et al. 2010; Meuwissen et al. 2013). Previous studies have suggested that more than half of the germplasm in six-row barley cultivars released in the Midwest U.S. after 1971 was contributed by five ancestral cultivars and the entire gene pool of North American six-row malt barley can be traced to 11 ancestral cultivars (Martin et al. 1991). Furthermore, inbred recycling, a typical practice in advanced cycle breeding, leads to highly related inbreds and smaller effective population sizes (Condon et al. 2008). Combining the progeny from two breeding programs could increase *Ne* (Goddard 2009; Daetwyler et al. 2010; Meuwissen et al. 2013). This would explain why in our case the prediction accuracy using the M+N panel continued to increase with greater population size. We also found larger training panels resulted in smaller standard deviation around the mean suggesting that larger training panels could protect breeders from stochastic effects resulting in poor predictions.

Our random resampling experiment also indicated that even very small training panels (n = 50) could occasionally generate prediction accuracies higher than training population sizes of 750 (Fig 4). This indicates that, at least in our case, some larger training panels might be less informative than smaller training panels. If true, then simply increasing training population size *per se* may reduce prediction accuracy. Further work is needed to test this hypothesis and develop strategies to design small but accurate training populations. This would likely include ways to select lines for a training population based on genetic distance from the selection candidates in addition to other factors. We speculate that breeders could design smaller custom training panels from the pool of available resources to improve prediction accuracy. For example, using the *k*-means algorithm (Lorenz et al. 2012) or a relationship matrix among lines could be used to design such custom training panels.

To predict progeny from crosses between breeding programs (MxN), we found that the combined training panel (M+N) was more accurate than training panels comprised of lines from only one breeding program. Three major reasons for the better predictability of M+N over either M or N are: first, crossing of inbred lines between two breeding programs should break up linkage blocks existed within individual programs. Therefore, predictions based on LD are expected to be lower for MxN progeny when either M or N alone was used as training panel. Second, combined populations can capture across breeding program wide LD as opposed to capturing LD information within any single subpopulation alone. Third, the accuracy with which marker effects are estimated has a significant effect on predicting the breeding value of lines (Meuwissen et

al. 2013). Combined training panel with large population size can better estimate effect of alleles segregating in MxN progenies which otherwise may be segregating in low frequency or fixed in either M or N. For diverse breeding materials, high-density marker panels that ideally tag every gene were reported to improve predictability (Ober et al. 2012). In our case, we used only 340 markers and a larger marker panel may improve accuracy particularly when predicting a more diverse progeny population.

Because both the size and composition of the training panel affected prediction accuracy, we further investigated the combined effects of these factors. We found that pooling lines from M and N helped improve accuracy in across (i.e., M predicting NxN or vice versa) and among (i.e., M or N predicting MxN) breeding program predictions. In contrast, increasing the combined training panel sample size from 384 to 768 only resulted in modest increase in relative prediction accuracy. This suggested that changing training panel composition rather than increasing sample size had larger impact on improving or maintaining relative prediction accuracy (Fig. 3). Therefore, breeders interested in developing universal prediction models for collaborative breeding may gain more from careful design of training panels by pooling lines (and combining the genetic variation) from other breeding programs rather than simply increasing panel size.

Implications for breeding

The decision to implement GS to augment or replace phenotypic selection requires that GS deliver better gains per unit time and/or cost (Heffner et al. 2010; Technow et al. 2013). Assuming equal selection intensities, Technow et al. (2013) compared the merit of GS as a function of the selection cycle lengths using the equation $L_Y = (r_A/H_X) * L_X$;

where, L_Y and L_X are the cycle lengths of genomic and phenotypic selection respectively, H_X is the square root of the heritability of the target trait, and r_A is the genetic correlation between the target trait and indirect trait. They suggested that GS will be better than phenotypic selection when L_Y was shorter than $(r_A/H_X)*L_X$. In our barley breeding program, a typical phenotypic selection breeding cycle is 4 years, and a corresponding GS cycle is one year (or less). Therefore, because relative prediction accuracy (i.e., r_A/H_X) for all traits were greater than 0.25, GS is expected to be superior to phenotypic selection in terms of gain per unit time. When we initiated this study, the cost to genotype a line with 384 SNPs was about $20. Current costs with new technologies have reduced that cost substantially. In contrast, the cost of phenotyping remains the same or is increasing. We estimate our costs for phenotyping at one location for the traits of yield, FHB and DON to be $20, $5, $10, respectively. The heritability estimates that we used to generate relative prediction accuracies were based on multiple location trials. This suggests that in addition to an improvement in gain per unit time there should be as substantial increase in gain per unit cost by implementing GS.

Genomic selection may have a valuable role in collaborative breeding efforts and germplasm exchange among breeding programs. We found that the prediction accuracies were generally high when training panels and prediction candidates were from the same breeding program. In contrast, prediction accuracies were lower when a training panel from one program was used to predict lines from another. However, even low prediction accuracies could be useful when phenotypic data is difficult or expensive to generate. This is clearly the case with Fusarium head blight which requires specialized disease

screening nurseries and quality of data can vary dramatically from year to year. Lower accuracy models derived from more genetically distant training panels could be used initially at low selection intensities to cull the most poorly performing lines. Later, after more phenotypic data from the target-breeding program is available, models that are more accurate could be implemented for selection.

Table 1. Trait variance and heritability estimates for the three training panels (TP).

		Variance		
	TP*	**Line**	**Residual**	**H**
Plant height	M	20.09	9.35	0.85
	N	16.55	9.02	0.88
	M+N	18.42	9.13	0.87
Heading date	M	0.73	0.55	0.78
	N	2.44	0.57	0.94
	M+N	1.55	0.57	0.90
Grain yield	M	124,688	360,771	0.48
	N	49,946	159,684	0.55
	M+N	94,715	232,468	0.58
FHB	M	13.18	46.38	0.48
	N	21.30	47.41	0.57
	M+N	18.26	47.56	0.56
DON	M	22.86	70.28	0.49
	N	22.17	63.23	0.46
	M+N	22.44	69.93	0.48
Grain protein	M	0.3540	0.3487	0.72
	N	0.4492	0.1703	0.87
	M+N	0.3724	0.3011	0.76

* M = University of Minnesota (384 lines), N= North Dakota State University (384 lines)

Table 2. Trait variance and heritability estimates for the progeny sets derived from crosses among selected parents from the University of Minnesota and North Dakota State University breeding programs.

	Cross type*	Variance				H
		line	trial	Rep	Residual	
Grain yield	MxM	20.5	495.4		88.7	0.48
	MxN	15.9	410.9		90.7	0.41
	NxN	14.4	451.2		97.3	0.37
	Combined	18.6	452.2		92.4	0.45
Plant height	MxM	5.4	84.8		17.5	0.64
	MxN	13.7	83.0		22.8	0.78
	NxN	11.0	80.3		27.4	0.70
	Combined	10.1	82.7		22.6	0.73
Heading date	MxM	0.1	18.5		1.3	0.34
	MxN	1.7	18.8		2.1	0.85
	NxN	2.3	19.8		1.9	0.89
	Combined	1.3	18.8		2.0	0.82
FHB	MxM	1.5	4.8	6.8	15.6	0.36
	MxN	2.6	6.7	7.1	22.1	0.42
	NxN	0.8	4.9	9.1	32.7	0.13
	Combined	2.8	5.0	7.2	20.7	0.45
DON	MxM	2.1	70.9	3.6	31.0	0.35
	MxN	6.6	74.0	1.5	36.6	0.59
	NxN	6.2	74.3	3.9	39.5	0.56
	Combined	4.9	67.4	0.9	30.0	0.57
Grain Protein[¥]	Combined	0.6	1.7		0.6	0.83

* MxM = Crosses within University of Minnesota breeding germplasm (100 lines), NxN= Crosses within the North Dakota State University breeding germplasm (100 lines), MxN= Crosses among breeding germplasm (100 lines), Combined = All 300 progeny lines analyzed together.

[¥]Grain protein data was not included in the population composition study because protein data was not available for all ND parents. Hence, to avoid confounding from training panel sample size on prediction accuracy, grain protein data was not analyzed separately for three progeny types.

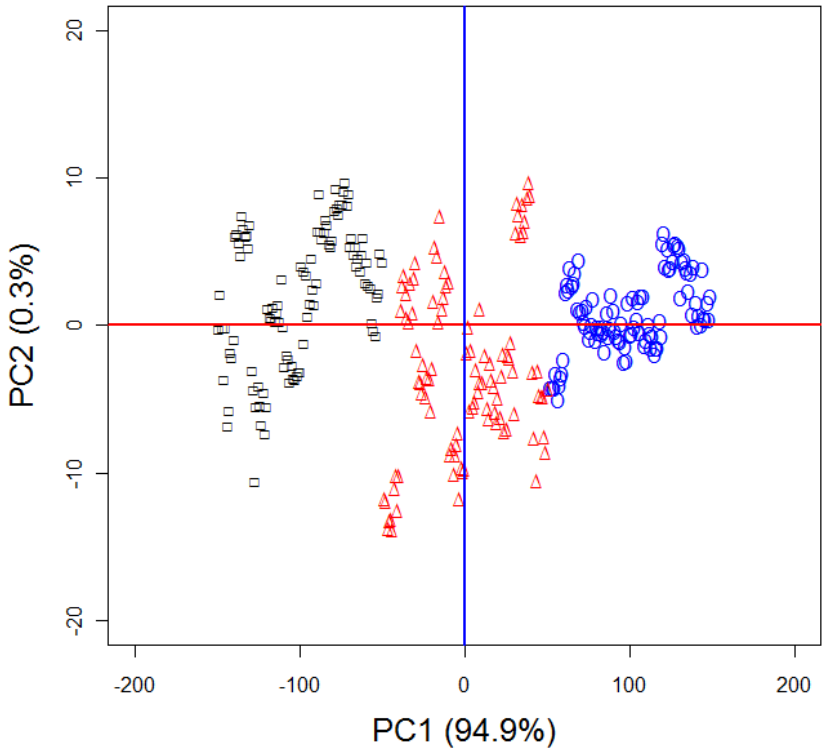

Fig.1 Plot of first (PC1) versus second (PC2) principal components for 300 progeny lines derived from Minnesota x Minnesota crosses (MxM; blue circles), North Dakota by North Dakota crosses (NxN; black squares) and Minnesota by North Dakota crosses (MxN; red triangles). Principal component analysis conducted using 340 SNP markers for each line.

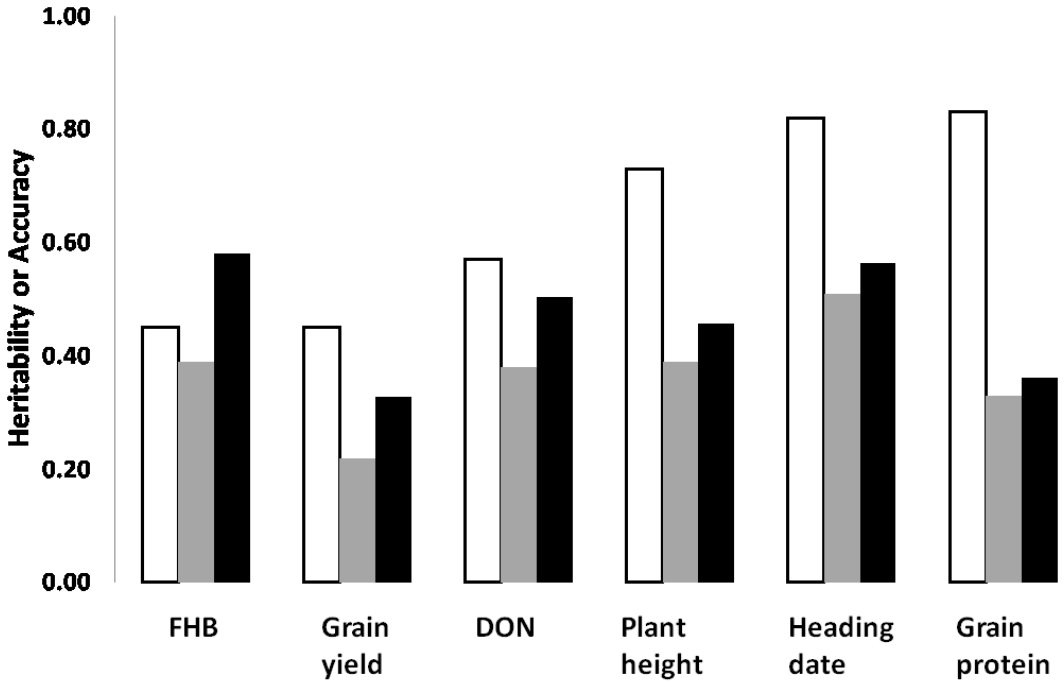

Fig.2 Heritability (white bar), prediction accuracy (grey bar), and relative prediction accuracy (black bar) for Fusarium head blight (FHB), yield, deoxynivalenol (DON), plant height, heading date, and grain protein. Marker-based predictions for the three hundred progeny lines were derived using RR BLUP and a training panel comprised of 768 lines from University of Minnesota and North Dakota State University breeding programs. Prediction accuracy was measured as the correlation between the marker-based prediction and phenotypic value. Relative prediction accuracy is the prediction accuracy divided by the square root of the heritability of trait.

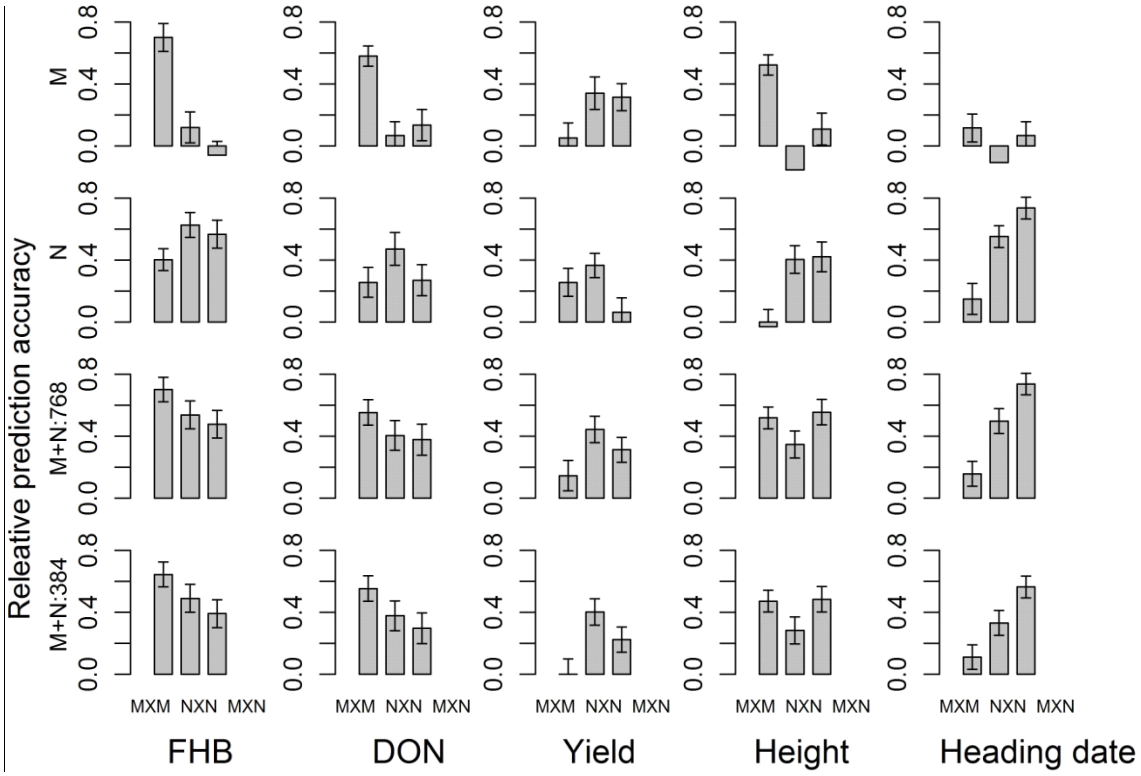

Fig.3 Comparison of relative prediction accuracy for four training panels from the University of Minnesota (M), North Dakota State University (N), both programs combined where n=768 (M+N:768) , and both programs combined where n=384 (M+N:384) predicting three cross types (MxM, NxN and MxN).

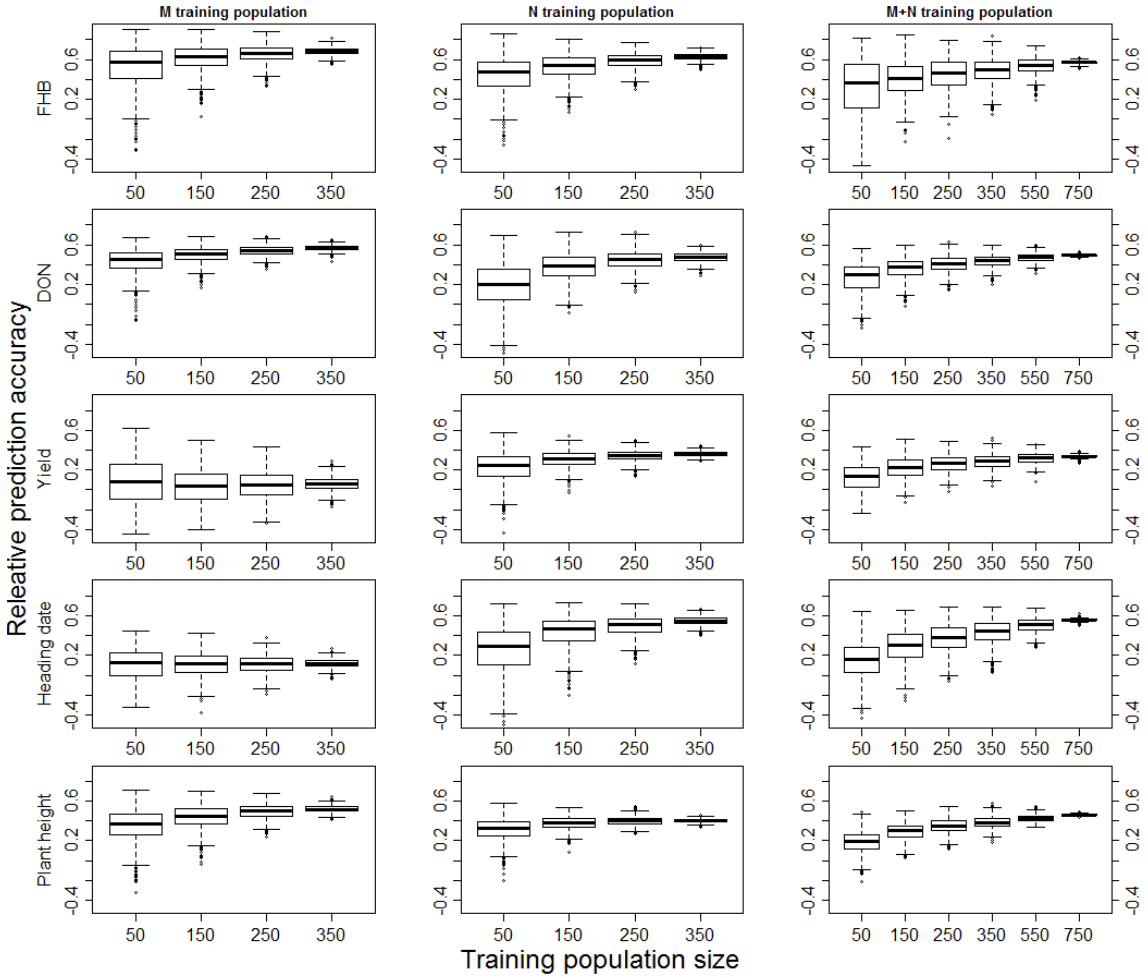

Fig.4 Effect of training population size on relative prediction accuracy using three training populations for the five traits Fusarium head blight severity (FHB), deoxynivalenol concentration (DON), yield, heading date, and plant height. For each sample size, one thousand training populations were randomly sampled. For each trait, the University of Minnesota training population (M) was used to predict 100 MxM progeny, North Dakota State University training population (N) was used to predict 100 NXN progeny, and both programs combined (M+N) where n=768 was used to predict all 300 progeny lines.

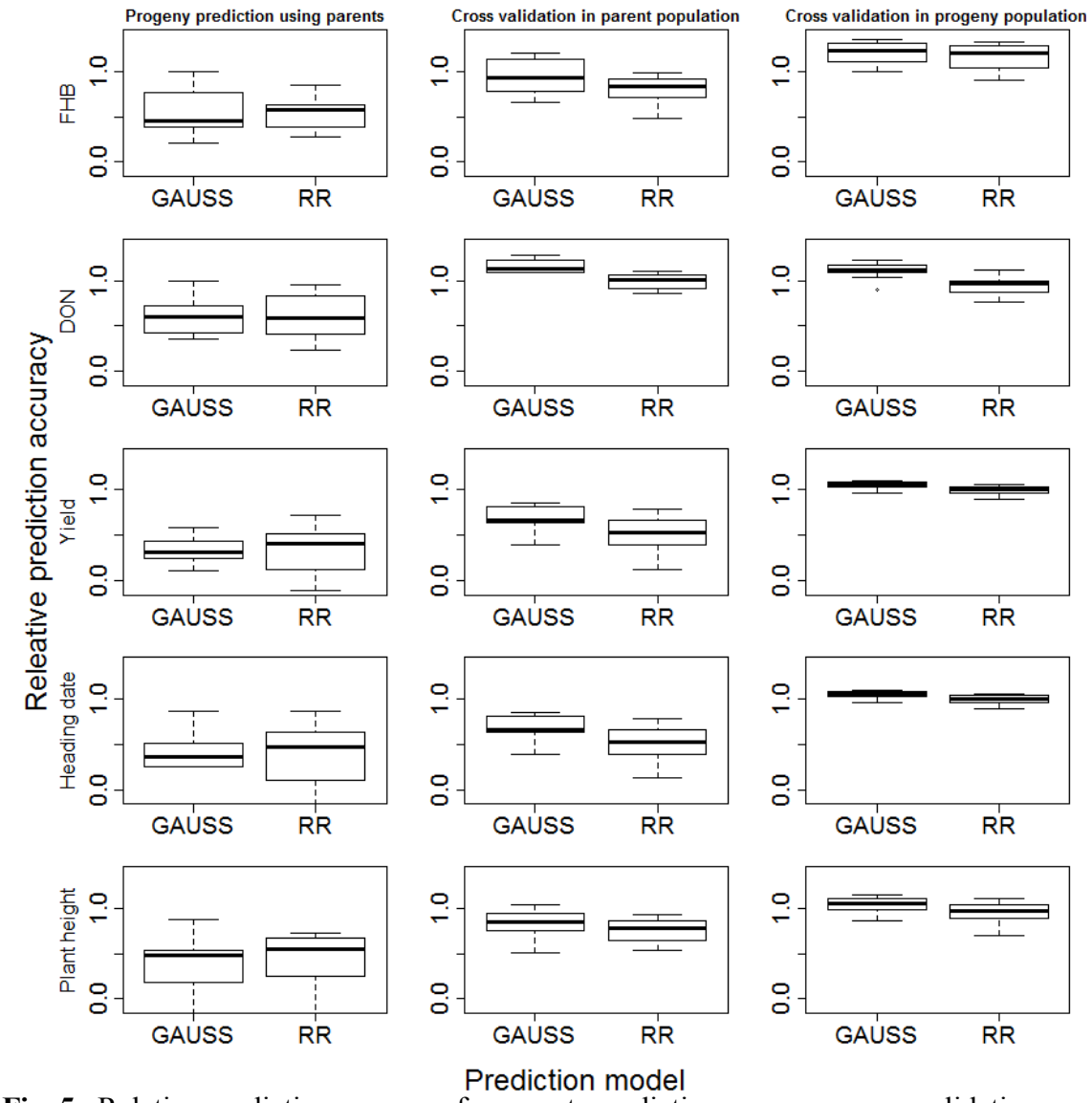

Fig. 5. Relative prediction accuracy for parents predicting progeny, cross-validation within the parent population, and cross-validation within the progeny population. Comparison between Gaussian kernel method (GAUSS) and random regression best linear unbiased predictor (RR). For each case, a training panel of 280 was used to predict a validation panel of 20. The relative prediction accuracy was estimated through a 10 round resampling procedure without replacement.

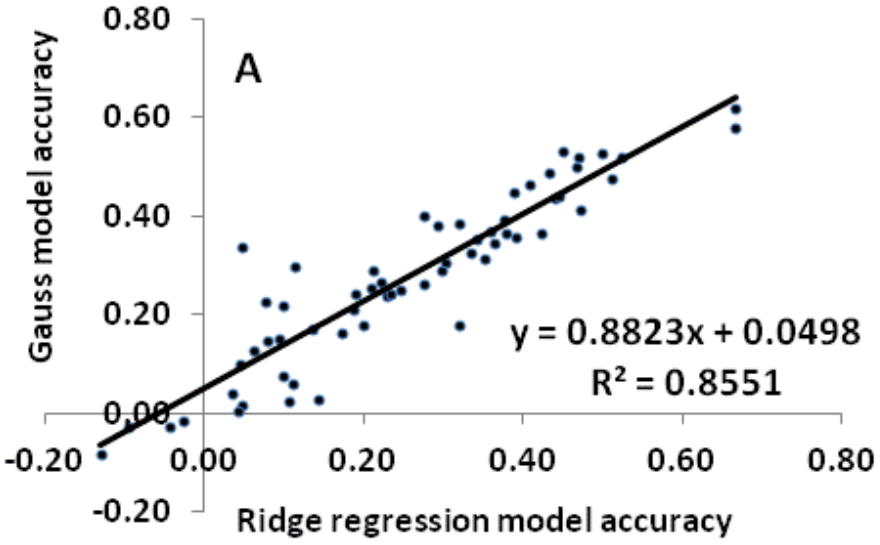

Fig. 6 Relationship between prediction accuracy estimated using RR BLUP model and gauss model for five traits and twelve training panel-progeny set combinations (n=60).

Reference

Asoro, F. G., Newell, M. A., Beavis, W. D., Scott, M. P., & Jannink, J. L. (2011). Accuracy and training population design for genomic selection on quantitative traits in elite North American oats. The Plant Genome, 4(2), 132-144.

Bernardo R (2008) Molecular markers and selection for complex traits in plants: learning from the last 20 years. Crop Sci 48:1649-1664

Bernardo R, Yu J (2007) Prospects for genomewide selection for quantitative traits in maize. Crop Sci 47:1082–1090

Blake V, Kling J, Hayes P, Jannink J-L, Jillella S, Lee J, Matthews D, Chao S, Close T, Muehlbauer G, Smith K, Wise R, Dickerson J (2012) The Hordeum toolbox—The barley CAP genotype and phenotype resource. Plant Genome 5:81–91

Close, T., Bhat, P., Lonardi, S., Wu, Y., Rostoks, N., Ramsay, L., Druka, A., & Waugh, R. (2009). Development and implementation of high-throughput SNP genotyping in barley. BMC genomics, 10(1), 582.

Condón, F., C. Gustus , D.C. Rasmusson, and K.P. Smith. 2008. Effect of Advanced Cycle Breeding on Genetic Diversity in Barley Breeding Germplasm. Crop Sci. 48:1027–1036.

Crossa, J., de Los Campos, G., Pérez, P., Gianola, D., Burgueño, J., Araus, J. L., ... & Braun, H. J. (2010). Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics, 186(2), 713-724.

Daetwyler, H. D., Pong-Wong, R., Villanueva, B., & Woolliams, J. A. (2010). The impact of genetic architecture on genome-wide evaluation methods. Genetics, 185(3), 1021-1031.

De Los Campos, G., Naya, H., Gianola, D., Crossa, J., Legarra, A., Manfredi, E., Weigel, K., & Cotes, J. M. (2009). Predicting quantitative traits with regression models for dense molecular markers and pedigree. Genetics, 182(1), 375-385.

Eathington SR, Crosbie TM, Edwards MD, Reiter RD, Bull JK (2007) Molecular markers in a commercial breeding program. Crop Sci 47: S154–S163

Efron, B., and G. Gong. 1983. A leisurely look at the bootstrap, the jackknife, and cross-validation. Am. Stat. 37: 36-48.

Endelman, J. B. (2011). Ridge regression and other kernels for genomic selection with R package rrBLUP. The Plant Genome, 4(3), 250-255.

Fan JB, Gunderson KL, Bibikova M, Yeakley JM, Chen J, Wickham Garcia E, Lebruska LL, Laurent M, Shen R, Barker D (2006) Illumina universal bead arrays. Methods Enzymol 410:57–73

Federer, W. T., & Raghavarao, D. (1975). On augmented designs. Biometrics, 29-35.

Gianola, D., & van Kaam, J. B. (2008). Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits. Genetics, 178(4), 2289-2303.

Goddard, M. (2009). Genomic selection: prediction of accuracy and maximisation of long term response. Genetica, 136(2), 245-257.

Hayes, B. J., Pryce, J., Chamberlain, A. J., Bowman, P. J., & Goddard, M. E. (2010). Genetic architecture of complex traits and accuracy of genomic prediction: coat colour, milk-fat percentage, and type in Holstein cattle as contrasting model traits. PLoS Genetics, 6(9), e1001139.

Hayes, B.J., P.J. Bowman, A.C. Chamberlain, K. Verbyla and M.E. Goddard. 2009. Accuracy of genomic breeding values in multi-breed dairy cattle populations. Genetics, Selection, Evolution : GSE 41:51.

Heffner, E. L., Jannink, J. L., & Sorrells, M. E. (2011). Genomic selection accuracy using multifamily prediction models in a wheat breeding program. The Plant Genome, 4(1), 65-75.

Heffner, E.L., A.J.Lorenz, J.-L.Jannink and M.E.Sorrells (2010) Plant breeding with genomic selection: gain per unit time and cost. Crop Sci. 50: 1681–1690.

Hofheinz, N., Borchardt, D., Weissleder, K., & Frisch, M. (2012). Genome-based prediction of test cross performance in two subsequent breeding cycles. Theoretical and Applied Genetics, 125(8), 1639-1645.

Jones, R. K., & Mirocha, C. J. (1999). Quality parameters in small grains from Minnesota affected by Fusarium head blight. Plant disease, 83(6), 506-511.

Legarra A, Robert-Grani_e C, Manfredi E, Elsen JM (2008) Performance of genomic selection in mice. Genetics 180: 611–618. doi: 10.1534/genetics.108.088575.

Li, H., Ribaut, J. M., Li, Z., & Wang, J. (2008). Inclusive composite interval mapping (ICIM) for digenic epistasis of quantitative traits in biparental populations. Theoretical and Applied Genetics, 116(2), 243-260.

Lorenz AJ, Chao S, Asoro FG, Heffner EL, Hayashi T, et al. (2011) Genomic selection in plant breeding: Knowledge and prospects. Adv Agron 110: 77–123.

Lorenz, A. J., Smith, K. P., & Jannink, J. L. (2012). Potential and optimization of genomic selection for Fusarium head blight resistance in six-row barley. Crop Science, 52(4), 1609-1621.

Lorenzana, R. E., & Bernardo, R. (2009). Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. Theoretical and applied genetics, 120(1), 151-161.

Martin JM, Blake TK, Hockett EA (1991) Diversity among North American spring barley cultivars based on coefficients of parentage. Crop Sci 31:1131-1137

Massman J, Cooper B, Horsley R, Neate S, Dill-Macky R, Chao S, Dong Y, Schwarz P, Muehlbauer GJ, Smith KP (2011) Genome-wide association mapping of Fusarium head blight resistance in contemporary barley breeding germplasm. Mol Breed doi: 10.1007/s11032-010-9442-0

Massman, J. M., Jung, H. J. G., & Bernardo, R. (2013). Genomewide selection versus marker-assisted recurrent selection to improve grain yield and stover-quality traits for cellulosic ethanol in maize. Crop Science, 53(1), 58-66.

Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genomewide dense marker maps. Genetics 157: 1819–1829.

Meuwissen, T., Hayes, B., & Goddard, M. (2013). Accelerating Improvement of Livestock with Genomic Selection. Annu. Rev. Anim. Biosci., 1(1), 221-237.

Ober, U., Erbe, M., Long, N., Porcu, E., Schlather, M., & Simianer, H. (2011). Predicting genetic values: a kernel-based best linear unbiased prediction with genomic data. Genetics, 188(3), 695-708.

Peel, M. D., & Rasmusson, D. C. (2000). Improvement strategy for mature plant breeding programs. Crop science, 40(5), 1241-1246.

Rasmusson DC, Phillips RL. Plant breeding progress and genetic diversity from de novo variation and elevated epistasis. Crop Sci. 1997;37(2):303–310.

Slotta TAB, Brady L, Chao S (2008) High throughput tissue preparation for large scale genotyping experiments. Mol Ecol Res 8

Technow, F., Bürger, A., & Melchinger, A. E. (2013). Genomic prediction of northern corn leaf blight resistance in maize with combined or separated training sets for heterotic groups. G3: Genes| Genomes| Genetics, 3(2), 197-203.

Toosi, A., Fernando, R. L., & Dekkers, J. C. M. (2010). Genomic selection in admixed and crossbred populations. Journal of animal science, 88(1), 32-46.

Whittaker JC, Thompson R, Denham MC Marker-assisted selection using ridge regression. Genetical research 2000;75:249-52

Xu, S., & Jia, Z. (2007). Genomewide analysis of epistatic effects for quantitative traits in barley. Genetics, 175(4), 1955-1963.

Zhong, S., Dekkers, J. C., Fernando, R. L., & Jannink, J. L. (2009). Factors affecting accuracy from genomic selection in populations derived from multiple inbred lines: a barley case study. Genetics, 182(1), 355-364.

Chapter 3

Future directions

The goal of my PhD dissertation was to explore novel ways to improve mature breeding programs that was genetically narrow and genetically distant from diverse unimproved germplasm. We coupled GWAS and GS to compare genetic architecture of traits in breeding programs, and explore the possibilities of employing genomic prediction approaches to design sound breeding and improvement strategies. In general, we found that different sets of QTL were segregating in the two breeding programs. Through informed germplasm exchange, both programs should be able to integrate novel alleles segregating in elite backgrounds without large linkage drag issues. However, further research is warranted on the effect of linkage drag. We are also confident that genomic selection approach holds excellent potential over phenotypic selection to improve breeding efficiency of yield and FHB that show complex inheritance and low to moderate heritability. Given these results, listed below are some potential follow-up projects for my study:

1. Long LD blocks in breeding germplasm were a major limiting factor that affected the resolution of QTL mapping. If the objective is identification of causal loci, we need to increase recombination within our breeding lines. Carefully designed highly recombined synthetic populations as in maize NAM (Yu et al. 2008) or Arabidopsis MAGIC (Kover at al. 2009) could be implemented in six-row malt barley to achieve these goals. In addition, to developing synthetic populations,

high density genotyping, preferably GBS or whole genome resequencing could be implemented to increase marker coverage, and avoid ascertainment bias. A recent GWAS study using drosophila MAGIC populations demonstrated the potentials for identifying allelic variants within a QTL region (King et al. 2012). Because plant breeders can generate immortal RIL that could be tested extensively in different environments, these multiparent populations, could enable breeders to identify allelic variants and detect allele by environment interactions. I speculate these results may aid breeders to identify and integrate best alleles particularly for traits that show high genotype by environment interaction e.g., cold tolerance or draught tolerance.

2. Although GS was found promising to improve complex agronomic traits, breeders should continue to identify ways to optimize and improve predictions accuracies. I found small training populations of size 50 could occasionally generate higher prediction accuracies than our largest training population. Although exploring that observation was beyond the scope of our study, potentials of developing custom training populations for a given prediction population warrants further research. I speculate that breeders could design smaller custom training panels from the pool of available resources to improve prediction accuracy. For example, k-means algorithm (Lorenz et al. 2012) or relationship matrix among lines could be used to design such custom training panels.

3. Because I have only tested the prediction accuracy of lines in cycle1, an interesting question that remains is to assess whether prediction accuracies decay in successive cycles. Although this scenario has been previously reported in animal breeding, their population histories very different from what is typically used in plant breeding. Additional research is needed to explore the model accuracy in later generations. If needed, breeders should consider retraining the models to improve/maintain prediction accuracy.
4. Although GS can greatly reduce breeding cycle and increase genetic gain per unit time, ways to further reduce the breeding cycles should be explored. Incorporating seed chipping technology and doubled haploid technology to GS are potential options to achieve these goals.
5. Because the accuracy of genomic prediction equations developed from one breeding program poorly predicted performance of lines in other breeding program, breeders might be skeptical to implement GS in their program using data from other programs. On the other hand, our results also indicated that the cross validation predictions accuracies could be greatly over inflated due to common genetic covariance among lines and sharing of common test environments among training and validation population. Although inflation of accuracy in cross validation is discouraging for breeders planning to speculate GS breeding efficiency, it could be used for breeder's benefit. For example, if a breeder wants to test 10,000 lines for a trait but he has no historical data to develop genomic prediction models or is skeptical of using models developed

from other breeders. In this scenario, breeder could plan to genotype and phenotype a smaller subset (~500 lines), develop prediction models, and use this model to predict rest (9,950). Based on our cross validation result, I speculate that the prediction accuracy based on models developed from the above 500 lines will be higher than models developed from other breeding programs due to shared genetic covariance between training and prediction populations. However, improved accuracy will be achieved at the expense of time. Therefore, more research is needed to compare the genetic gain for above two-step approach vs leveraging data from other breeders.

6. Possibilities and potentials of combining above described synthetic populations development with genomic selection approaches needs to be explored. QTL mapping in MAGIC population has the potential to identify lines with superior alleles. These lines could be used as parents to introgress genes of interest via genomic selection.
7. Strong directional selection has the potential to generate large selective sweeps around favorable alleles in the genome. Because GS aims to capture total additive genetic variance using many genome wide markers, it can also result in large genome wide selective sweeps. This scenario along with fast breeding cycle time and high selection intensity may rapidly fix many markers in subsequent generations and increase coefficient of coancestry among lines in the advanced cycles of selection. I speculate that the situation may be more detrimental if low-density marker panels that span the whole genome are used to conduct repeated

cycles of selection. For unit time, the gains may continually diminish at a faster rate in genomic selection than phenotypic selection. Furthermore, because most breeders work with small populations, genetic drift can also play a big role in limiting the gains through loss of alleles (and genetic diversity). The major topics that need further investigation are: 1) Investigate impact of recurrent genomic selection on genetic variance and gain from selection through a multigenerational selection experiments. 2) Compare the efficiency of genomic selection over classical phenotypic selection using multigenerational selection experiments, and 3) Determine the impact of selective sweep and drift on gain from genomic selection. Model organisms with very short generation time could be useful to test these hypotheses.

Bibliography

Asoro, F. G., Newell, M. A., Beavis, W. D., Scott, M. P., & Jannink, J. L. (2011). Accuracy and training population design for genomic selection on quantitative traits in elite North American oats. The Plant Genome, 4(2), 132-144.

Beavis, W.D., 1998. QTL analyses: power, precision, and accuracy, pp. 145–162 in Molecular Dissection of Complex Traits, edited by A.H. Paterson. CRC Press, Boca Raton,

Belknap, J. K., Mitchell, S. R., O'Toole, L. A., Helms, M. L., & Crabbe, J. C. (1996). Type I and type II error rates for quantitative trait loci (QTL) mapping studies using recombinant inbred mouse strains. Behavior genetics, 26(2), 149-160.

Bernardo R (2008) Molecular markers and selection for complex traits in plants: learning from the last 20 years. Crop Sci 48:1649-1664

Bernardo R, Yu J (2007) Prospects for genomewide selection for quantitative traits in maize. Crop Sci 47:1082–1090

Blake V, Kling J, Hayes P, Jannink J-L, Jillella S, Lee J, Matthews D, Chao S, Close T, Muehlbauer G, Smith K, Wise R, Dickerson J (2012) The Hordeum toolbox—The barley CAP genotype and phenotype resource. Plant Genome 5:81–91

Bradbury P, Parker T, Hamblin MT, Jannink JL (2011) Assessment of power and false discovery in genome-wide association studies using the BarleyCAP germplasm. Crop Sci 51:52-59

Breseghello, F., & Sorrells, M. E. (2006). Association mapping of kernel size and milling quality in wheat (Triticum aestivum L.) cultivars. Genetics, 172(2), 1165-1177.

Bush WS, Moore JH (2012) Chapter 11: Genome-Wide Association Studies. PLoS Comput Biol 8(12): e1002822. doi:10.1371/journal.pcbi.1002822

Canci, P. C., Nduulu, L. M., Muehlbauer, G. J., Dill-Macky, R., Rasmusson, D. C., & Smith, K. P. (2004). Validation of quantitative trait loci for Fusarium head blight and kernel discoloration in barley. Molecular Breeding, 14(2), 91-104.

Carlson, C. S., Eberle, M. A., Kruglyak, L., & Nickerson, D. A. (2004). Mapping complex disease loci in whole-genome association studies. Nature, 429(6990), 446-452.

Close TJ, Bhat PR, Lonardi S, Wu Y, Rostoks N, Ramsay L, Druka A, Stein N et al (2009) Development and imple-mentation of high-throughput SNP genotyping in barley. BMC Genomics 10:582

Close, T., Bhat, P., Lonardi, S., Wu, Y., Rostoks, N., Ramsay, L., Druka, A., & Waugh, R. (2009). Development and implementation of high-throughput SNP genotyping in barley. BMC genomics, 10(1), 582.

Comadran J, Ramsay L, MacKenzie K, Hayes P, Close TJ, Muehlbauer G, Stein N, Waugh R.2011. Theor Appl Genet. 2011 Feb;122(3):523-31.

Comadran, J., Russell, J. R., Van Eeuwijk, F. A., Ceccarelli, S., Grando, S., Baum, M., ... & Thomas, W. T. B. (2008). Mapping adaptation of barley to droughted environments. Euphytica, 161(1-2), 35-45.

Condón[a], F., C. Gustus , D.C. Rasmusson, and K.P. Smith. 2008. Effect of Advanced Cycle Breeding on Genetic Diversity in Barley Breeding Germplasm. Crop Sci. 48:1027–1036.

Condon[b], F., D.C. Rasmusson1, E. Schiefelbein G. Velasquez, and K.P. Smith. 2008. Effect of Advanced Cycle Breeding on Genetic Gain and Phenotypic Diversity in Barley Breeding Germplasm. Crop Sci. 49: 1751-1761.

Crossa, J., de Los Campos, G., Pérez, P., Gianola, D., Burgueño, J., Araus, J. L., ... & Braun, H. J. (2010). Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics, 186(2), 713-724.

Daetwyler, H. D., Pong-Wong, R., Villanueva, B., & Woolliams, J. A. (2010). The impact of genetic architecture on genome-wide evaluation methods. Genetics, 185(3), 1021-1031.

Dahleen, L. S., Agrama, H. A., Horsley, R. D., Steffenson, B. J., Schwarz, P. B., Mesfin, A., & Franckowiak, J. D. (2003). Identification of QTLs associated with Fusarium head blight resistance in Zhedar 2 barley. Theoretical and applied genetics, 108(1), 95-104.

De la Pena, R. C., Smith, K. P., Capettini, F., Muehlbauer, G. J., Gallo-Meagher, M., Dill-Macky, R., ... & Rasmusson, D. C. (1999). Quantitative trait loci associated with resistance to Fusarium head blight and kernel discoloration in barley. Theoretical and Applied Genetics, 99(3-4), 561-569.

De Los Campos, G., Naya, H., Gianola, D., Crossa, J., Legarra, A., Manfredi, E., Weigel, K., & Cotes, J. M. (2009). Predicting quantitative traits with regression models for dense molecular markers and pedigree. Genetics, 182(1), 375-385.

Dorweiler, J., Stec, A., Kermicle, J., & Doebley, J. (1993). Teosinte glume architecture 1: a genetic locus controlling a key step in maize evolution. Science, 262(5131), 233-235.

Druka, A., Franckowiak, J., Lundqvist, U., Bonar, N., Alexander, J., Houston, K., ... & Waugh, R. (2011). Genetic dissection of barley morphology and development. Plant physiology, 155(2), 617-627.

Eathington SR, Crosbie TM, Edwards MD, Reiter RD, Bull JK (2007) Molecular markers in a commercial breeding program. Crop Sci 47: S154–S163

Efron, B., and G. Gong. 1983. A leisurely look at the bootstrap, the jackknife, and cross-validation. Am. Stat. 37: 36-48.

Endelman, J. B. (2011). Ridge regression and other kernels for genomic selection with R package rrBLUP. The Plant Genome, 4(3), 250-255.

Fan JB, Gunderson KL, Bibikova M, Yeakley JM, Chen J, Wickham Garcia E, Lebruska LL, Laurent M, Shen R, Barker D (2006) Illumina universal bead arrays. Methods Enzymol 410:57–73

Federer, W. T., & Raghavarao, D. (1975). On augmented designs. Biometrics, 29-35.

Feuillet C, Langridge P, Waugh R. Cereal breeding takes a walk on the wild side. Trends in Genetics 2008;24:24-32.

Gebhardt, D.J., D.C. Rasmusson, and R.D. Wilcoxson. Cyclic breeding to incorporate kernel discoloration resistance into malting barley. Crop Sci. 1992. 32:352–356.

Gianola, D., & van Kaam, J. B. (2008). Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits. Genetics, 178(4), 2289-2303.

Goblirsch, C.A., R.D. Horsley, and P.B. Schwarz. A strategy to breed low-protein barley with acceptable kernel color and diastatic power. Crop Sci. 1996. 36:41–44

Goddard, M. (2009). Genomic selection: prediction of accuracy and maximisation of long term response. Genetica, 136(2), 245-257.

Hajjar, R., & Hodgkin, T. (2007). The use of wild relatives in crop improvement: a survey of developments over the last 20 years. Euphytica, 156(1-2), 1-13.

Hayes, B. J., & Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. Genetics, 157(4), 1819-1829.

Hayes, B. J., Pryce, J., Chamberlain, A. J., Bowman, P. J., & Goddard, M. E. (2010). Genetic architecture of complex traits and accuracy of genomic prediction: coat colour, milk-fat percentage, and type in Holstein cattle as contrasting model traits. PLoS Genetics, 6(9), e1001139.

Hayes, B.J., P.J. Bowman, A.C. Chamberlain, K. Verbyla and M.E. Goddard. 2009. Accuracy of genomic breeding values in multi-breed dairy cattle populations. Genetics, Selection, Evolution : GSE 41:51.

Heffner, E. L., Jannink, J. L., & Sorrells, M. E. (2011). Genomic selection accuracy using multifamily prediction models in a wheat breeding program. The Plant Genome, 4(1), 65-75.

Hofheinz, N., Borchardt, D., Weissleder, K., & Frisch, M. (2012). Genome-based prediction of test cross performance in two subsequent breeding cycles. Theoretical and Applied Genetics, 125(8), 1639-1645.

Horsley, R. D., Schmierer, D., Maier, C., Kudrna, D., Urrea, C. A., Steffenson, B. J., ... & Kleinhofs, A. (2006). Identification of QTLs associated with Fusarium head blight resistance in barley accession CIho 4196. Crop science, 46(1), 145-156.

Horsley, R. D., Schwarz, P. B., & Hammond, J. J. (1995). Genetic diversity in malt quality of North American six-rowed spring barley. Crop science, 35(1), 113-118.

Huang Y, Millett BP, Beaubien KA, Dahl SK, Steffenson BJ, Smith KP, Muehlbauer GJ. (2012). Haplotype diversity and population structure in cultivated and wild barley evaluated for Fusarium head blight responses. Theor. Appl. Genet. 126 (3): 619-36.

Ioannidis, J.P., Thomas, G. & Daly, M.J. Validating, augmenting and refining genome-wide association signals. Nat. Rev. Genet. 10, 318–329 (2009).

Jannink, J. L., & Walsh, B. (2002). Association mapping in plant populations. Quantitative genetics, genomics and plant breeding, 59-68.

Jones, R. K., & Mirocha, C. J. (1999). Quality parameters in small grains from Minnesota affected by Fusarium head blight. Plant disease, 83(6), 506-511.

Kang et al. (2008) Efficient control of population structure in model organism association mapping. Genetics. 2008 Mar;178(3):1709-23.

King, E. G., C. M. Merkes, C. L. McNeil, S. R. Hoofer, S. Senet al.,2012 Genetic dissection of a model complex trait using the Drosophila Synthetic Population Resource. Genome Res. 22

Kover, P. X., Valdar, W., Trakalo, J., Scarcelli, N., Ehrenreich, I. M., Purugganan, M. D., ... & Mott, R. (2009). A multiparent advanced generation inter-cross to fine-map quantitative traits in Arabidopsis thaliana. PLoS genetics, 5(7), e1000551.

Lander E.S., and Kruglyak L. (1995) Genetic dissection of complex traits: Guidelines for interpreting and reporting linkage results. Nat. Genet. 11:241–247.

Legarra A, Robert-Grani_e C, Manfredi E, Elsen JM (2008) Performance of genomic selection in mice. Genetics 180: 611–618. doi: 10.1534/genetics.108.088575.

Li, H., Ribaut, J. M., Li, Z., & Wang, J. (2008). Inclusive composite interval mapping (ICIM) for digenic epistasis of quantitative traits in biparental populations. Theoretical and Applied Genetics, 116(2), 243-260.

Lin X.H., Zhang D.P., Xie Y.F., Gao H.P., Zhang, Q., 1996. Identifying and mapping a new gene for bacterial blight resistance in rice based on RFLP markers. Phytopathology 86: 1156-1159.

Long AD, Langley CH. 1999. The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits. Genome Res. 1999 Aug;9(8):720-31.

Lorenz AJ, Chao S, Asoro FG, Heffner EL, Hayashi T, et al. (2011) Genomic selection in plant breeding: Knowledge and prospects. Adv Agron 110: 77–123.

Lorenz, A. J., Smith, K. P., & Jannink, J. L. (2012). Potential and optimization of genomic selection for Fusarium head blight resistance in six-row barley. Crop Science, 52(4), 1609-1621.

Lorenzana, R. E., & Bernardo, R. (2009). Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. Theoretical and applied genetics, 120(1), 151-161.

Lundqvist, U. & Lundqvist, A. Induced intermedium mutants in barley: origin, morphology and inheritance. Hereditas 108, 13–26 (1988).

Martin JM, Blake TK, Hockett EA (1991) Diversity among North American spring barley cultivars based on coefficients of parentage. Crop Sci 31:1131-1137

Massman J, Cooper B, Horsley R, Neate S, Dill-Macky R, Chao S, Dong Y, Schwarz P, Muehlbauer GJ, Smith KP (2011) Genome-wide association mapping of Fusarium head blight resistance in contemporary barley breeding germplasm. Mol Breed doi: 10.1007/s11032-010-9442-0

Massman, J. M., Jung, H. J. G., & Bernardo, R. (2013). Genomewide selection versus marker-assisted recurrent selection to improve grain yield and stover-quality traits for cellulosic ethanol in maize. Crop Science, 53(1), 58-66.

Mesfin, A., Smith, K. P., Dill-Macky, R., Evans, C. K., Waugh, R., Gustus, C. D., & Muehlbauer, G. J. (2003). Quantitative trait loci for Fusarium head blight resistance in barley detected in a two-rowed by six-rowed population. Crop science, 43(1), 307-318.

Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genomewide dense marker maps. Genetics 157: 1819–1829.

Meuwissen, T., Hayes, B., & Goddard, M. (2013). Accelerating Improvement of Livestock with Genomic Selection. Annu. Rev. Anim. Biosci., 1(1), 221-237.

Nduulu, L. M., Mesfin, A., Muehlbauer, G. J., & Smith, K. P. (2007). Analysis of the chromosome 2 (2H) region of barley associated with the correlated traits Fusarium head blight resistance and heading date. Theoretical and Applied Genetics, 115(4), 561-570.

Ober, U., Erbe, M., Long, N., Porcu, E., Schlather, M., & Simianer, H. (2011). Predicting genetic values: a kernel-based best linear unbiased prediction with genomic data. Genetics, 188(3), 695-708.

Peel, M. D., & Rasmusson, D. C. (2000). Improvement strategy for mature plant breeding programs. Crop science, 40(5), 1241-1246.

Pillen-K, Zacharias-A, Leon-J, "Advanced backcross QTL analysis in barley (Hordeum vulgare L.)", Theor Appl Genet, 2003, vol. 107, pp. 340-352

Pritchard JK, Cox NJ. 2002. The allelic architecture of human disease genes: Common disease–common variant…or not? *Hum Mol Genet* 11*:* 2417–2423.

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. Genetics, 155(2), 945-959.

Pumphrey, M. O., Bernardo, R., & Anderson, J. A. (2007). Validating the QTL for Fusarium head blight resistance in near-isogenic wheat lines developed from breeding populations. Crop Science, 47(1), 200-206.

Rafalski, A. (2002). Applications of single nucleotide polymorphisms in crop genetics. Current opinion in plant biology, 5(2), 94-100.

Ramsay, L., Comadran, J., Druka, A., Marshall, D. F., Thomas, W. T., Macaulay, M., ... & Waugh, R. (2011). INTERMEDIUM-C, a modifier of lateral spikelet fertility in barley, is an ortholog of the maize domestication gene TEOSINTE BRANCHED 1. Nature genetics, 43(2), 169-172.

Rasmusson DC, Phillips RL. Plant breeding progress and genetic diversity from de novo variation and elevated epistasis. Crop Sci. 1997;37(2):303–310.

Rostoks N, Mudie S, Cardle L, Russell J, Ramsay L, Booth A, Svens-son JT, Wanamaker SI, Walia H, Rodriguez EM, Hedley PE, Liu H, Morris J, Close TJ, Marshall DF, Waugh R (2005) Genome-wide SNP

discovery and linkage analysis in barley based on genes responsive to abiotic stress. Mol Gen Genet 274:515–527

Scho¨n CC, Utz HF, Groh S, Truberg B, Openshaw S, Melchinger AE (2004) Quantitative trait locus mapping based on resampling in a vast maize testcross experiment and its relevance to quantitative genetics for complex traits. Genetics 167:485–498

Slotta TAB, Brady L, Chao S (2008) High throughput tissue preparation for large scale genotyping experiments. Mol Ecol Res 8

Smith, K. P., Budde, A., Dill-Macky, R., Rasmusson, D. C., Schiefelbein, E., Steffenson, B., Wiersma, J. J., Wiersma, J. V., & Zhang, B. (2013). Registration of 'Quest'spring malting barley with improved resistance to Fusarium head blight. Journal of Plant Registrations, 7(2), 125-129.

Technow, F., Bürger, A., & Melchinger, A. E. (2013). Genomic prediction of northern corn leaf blight resistance in maize with combined or separated training sets for heterotic groups. G3: Genes| Genomes| Genetics, 3(2), 197-203.

Tester, M., & Langridge, P. (2010). Breeding technologies to increase crop production in a changing world. Science, 327(5967), 818-822.

Toosi, A., Fernando, R. L., & Dekkers, J. C. M. (2010). Genomic selection in admixed and crossbred populations. Journal of animal science, 88(1), 32-46.

van Berloo, R., Aalbers, H., Werkman, A., & Niks, R. E. (2001). Resistance QTL confirmed through development of QTL-NILs for barley leaf rust resistance. Molecular Breeding, 8(3), 187-195.

Wang H, Smith KP, Combs E, Blake T, Horsley RD et al (2011) Effect of population size and unbalanced data sets on QTL detection using genome-wide association mapping in barley breeding germplasm. Theor Appl Genet

Weir, B. S., & Cockerham, C. C. (1984). Estimating F-statistics for the analysis of population structure. evolution, 1358-1370.

Whittaker JC, Thompson R, Denham MC Marker-assisted selection using ridge regression. Genetical research 2000;75:249-52

Wu¨rschum T, Maurer HP, Kraft T, Janssen G, Nilsson C et al (2011a) Genome-wide association mapping of agronomic traits in sugar beet. Theor Appl Genet 123:1121–1131

Wu¨rschum T, Maurer HP, Schulz B, Mo¨hring J, Reif JC (2011b) Genome-wide association mapping reveals epistasis and genetic interaction networks in sugar beet. Theor Appl Genet 123:109–118

Xu, S., & Jia, Z. (2007). Genomewide analysis of epistatic effects for quantitative traits in barley. Genetics, 175(4), 1955-1963.

Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Heath AC, Martin NG, Montgomery GW, et al. 2010. Common SNPs explain a large proportion of the heritability for human height. *Nat Genet* 42*:* 565–569.

Young, N. D. (1996). QTL mapping and quantitative disease resistance in plants. Annual review of phytopathology, 34(1), 479-501.

Young, N. D. (1999). A cautiously optimistic vision for marker-assisted breeding. Molecular breeding, 5(6), 505-510.

Yu, J., Pressoir, G., Briggs, W. H., Bi, I. V., Yamasaki, M., Doebley, J. F., ... & Buckler, E. S. (2005). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nature genetics, 38(2), 203-208.

Zhong, S., Dekkers, J. C., Fernando, R. L., & Jannink, J. L. (2009). Factors affecting accuracy from genomic selection in populations derived from multiple inbred lines: a barley case study. Genetics, 182(1), 355-364.

Zhou, H., & Steffenson, B. J. (2013). Genome-wide association mapping reveals genetic architecture of durable spot blotch resistance in US barley breeding germplasm. In Advance in Barley Sciences (pp. 257-267). Springer Netherlands.

Zhu, C et al. (2010) Status and Prospects of Association Mapping in Plants. The Plant Genome 1, 5-20.

Zhu, H., Gilchrist, L., Hayes, P., Kleinhofs, A., Kudrna, D., Liu, Z., Prom, L., Steffenson, B., Toojinda, T & Vivar, H. (1999). Does function follow form? Principal QTLs for Fusarium head blight (FHB) resistance are coincident with QTLs for inflorescence traits and plant height in a doubled-haploid population of barley. Theoretical and Applied Genetics, 99(7-8), 1221-1232.

Appendix

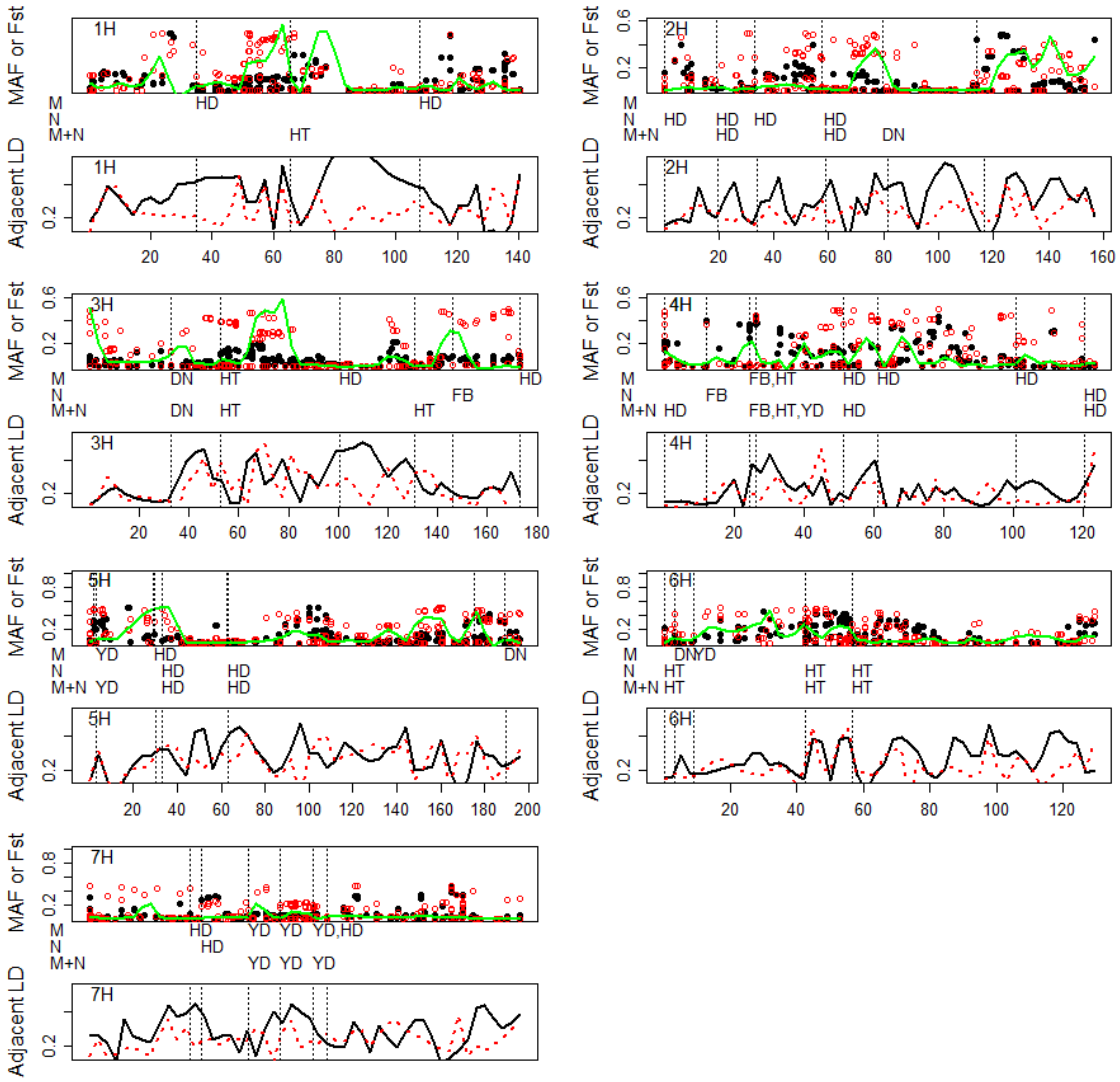

Fig.S1 Distribution of minor allele frequency (M = black solid circles, N = red open circles), adjacent marker linkage disequilibrium (M = black solid line, N = red broken line) and F_{ST} (green line) in the Minnesota and North Dakota breeding programs based on 2,110 single nucleotide polymorphisms. Vertical dotted lines indicate the position of a QTL which is labeled below as being detected in the Minnesota (M), North Dakota (N) or combined (M+N) mapping panels. Traits for QTL are labeled as Fusarium head blight severity (FH), deoxynivalenol concentration (DN), grain yield (YD), heading date (HD), and plant height (PH). Genetic distance (cM) is indicated above each chromosome panel.

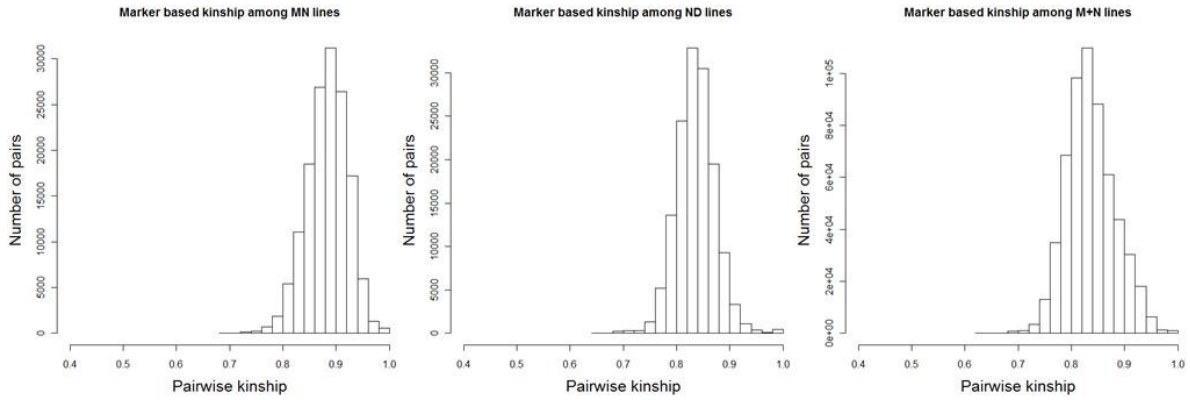

Fig.S2 Histogram of kinship coefficients estimated in M, N and M+N mapping panels from University of Minnesota and North Dakota State University breeding programs.

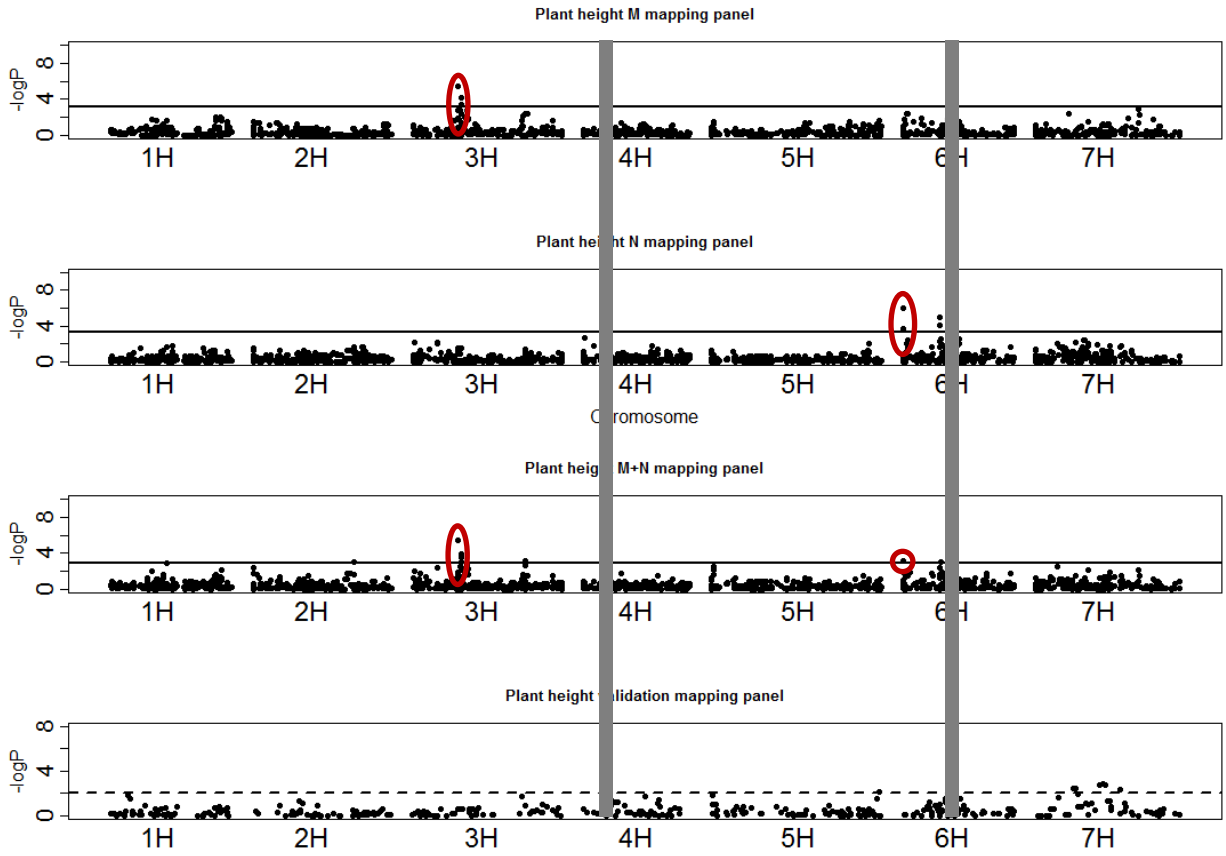

Fig.S3 Manhattan plot of p-values from association mapping for plant height in six-row spring barley germplasm. (A) Minnesota mapping panel (M), (B) North Dakota mapping panel (N), (C) Combined mapping panel (M+N), (D) Validation mapping panel. The solid line refers to the FDR cutoff of 1%. Because the validation mapping panels was genotyped only with 384 markers, the statistical cutoff (broken line) was suggestive (p = 0.01). Grey vertical bar indicates positions where a QTL was independently validated. Red circles indicate QTL detected in more than one mapping panel.

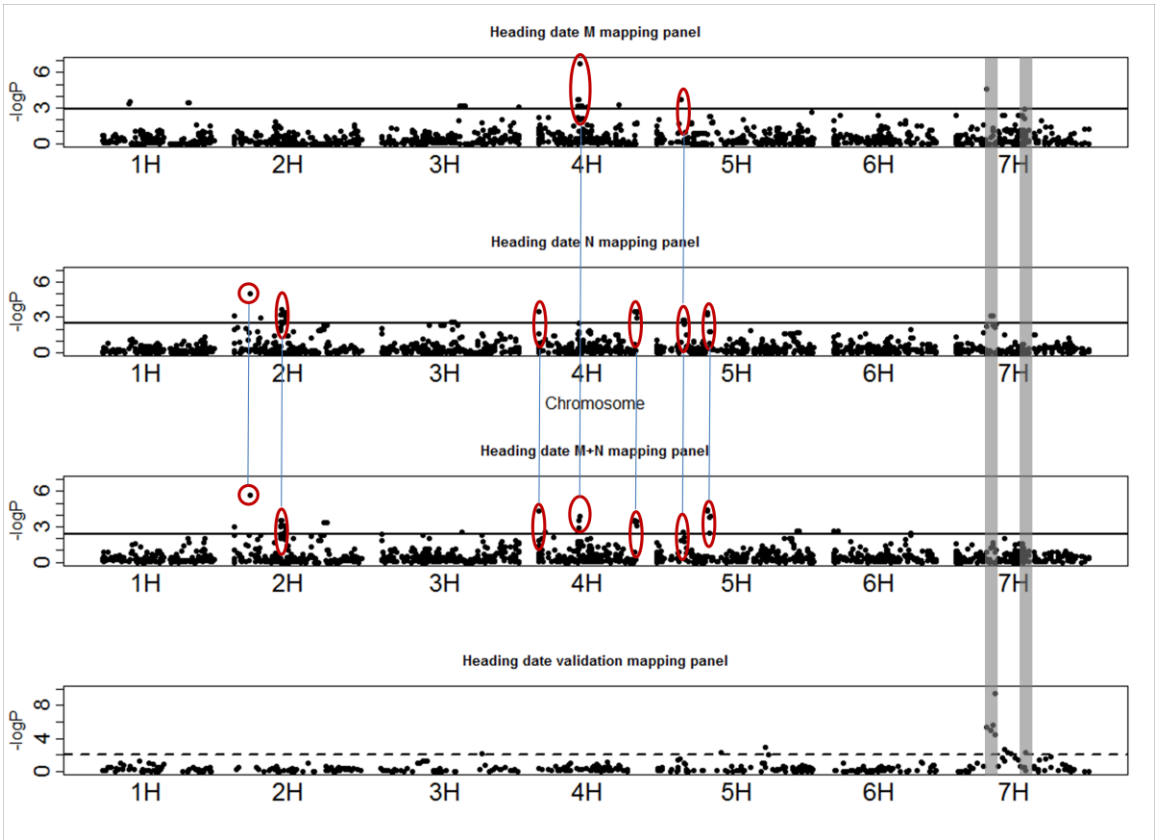

Fig.S4 Manhattan plot of p-values from association mapping for heading date in six-row spring barley germplasm. (A) Minnesota mapping panel (M), (B) North Dakota mapping panel (N), (C) Combined mapping panel (M+N), (D) Validation mapping panel. The solid line refers to the FDR cutoff of 1%. Because the validation mapping panels was genotyped only with 384 markers, the statistical cutoff (broken line) was suggestive (p = 0.01). Grey vertical bar indicates positions where a QTL was independently validated. Red circles indicate QTL detected in more than one mapping panel.

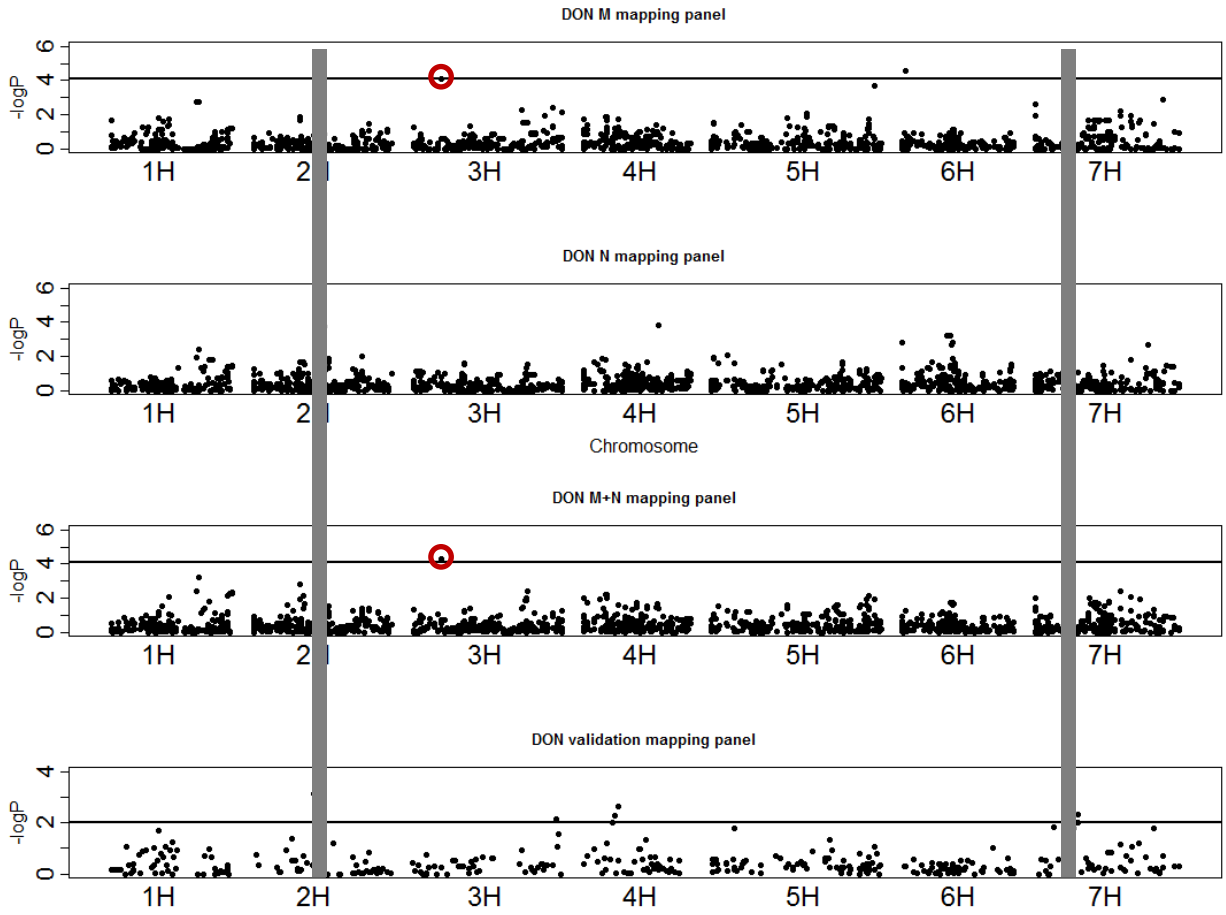

Fig.S5 Manhattan plot of p-values from association mapping for DON in six-row spring barley germplasm. (A) Minnesota mapping panel (M), (B) North Dakota mapping panel (N), (C) Combined mapping panel (M+N), (D) Validation mapping panel. The solid line refers to the FDR cutoff of 1%. Because the validation mapping panels was genotyped only with 384 markers, the statistical cutoff (broken line) was suggestive (p = 0.01). Grey vertical bar indicates positions where a QTL was independently validated. Red circles indicate QTL detected in more than one mapping panel.

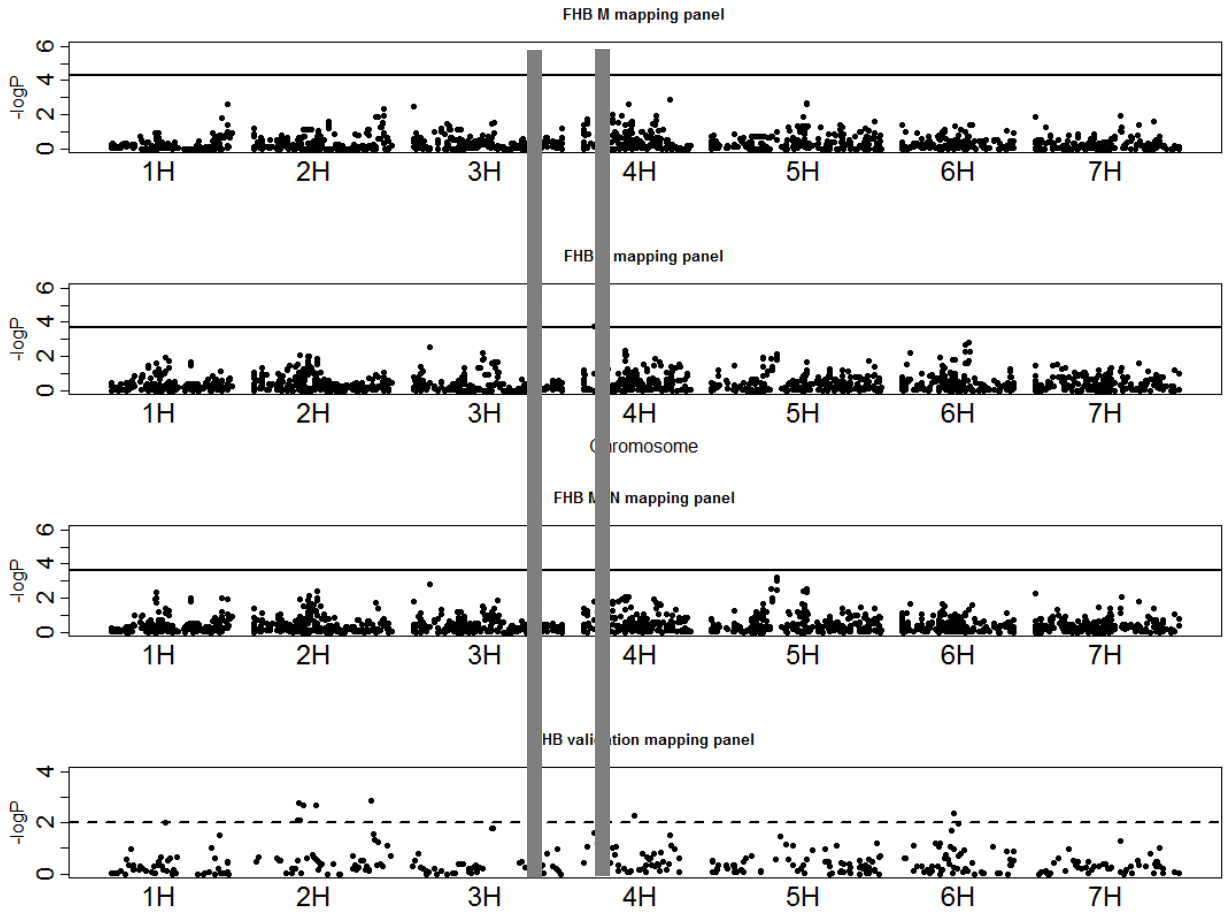

Fig.S6 Manhattan plot of p-values from association mapping for FHB in six-row spring barley germplasm. (A) Minnesota mapping panel (M), (B) North Dakota mapping panel (N), (C) Combined mapping panel (M+N), (D) Validation mapping panel. The solid line refers to the FDR cutoff of 1%. Because the validation mapping panels was genotyped only with 384 markers, the statistical cutoff (broken line) was suggestive (p = 0.01). Grey vertical bar indicates positions where a QTL was independently validated. Red circles indicate QTL detected in more than one mapping panel.

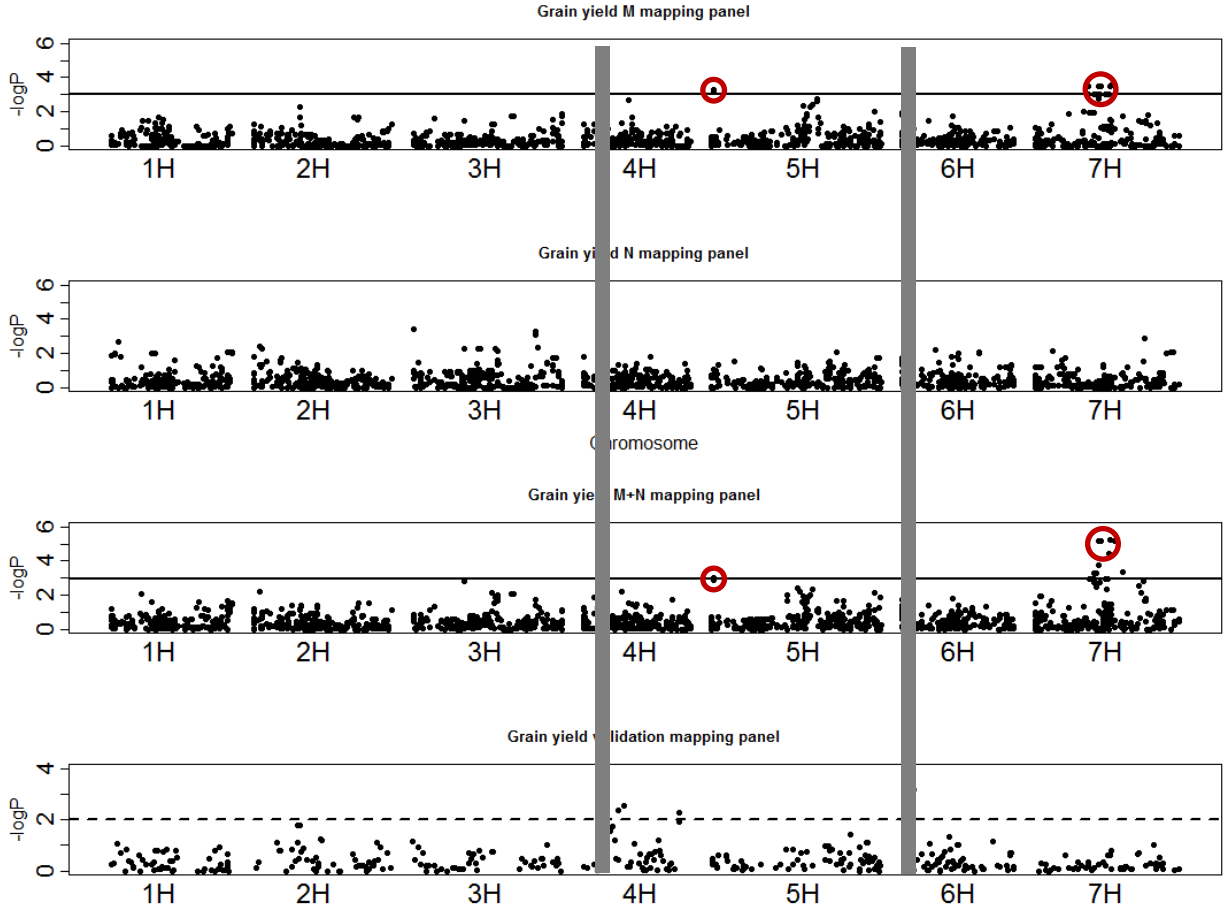

Fig.S7 Manhattan plot of p-values from association mapping for grain yield in six-row spring barley germplasm. (A) Minnesota mapping panel (M), (B) North Dakota mapping panel (N), (C) Combined mapping panel (M+N), (D) Validation mapping panel. The solid line refers to the FDR cutoff of 1%. Because the validation mapping panels was genotyped only with 384 markers, the statistical cutoff (broken line) was suggestive (p = 0.01). Grey vertical bar indicates positions where a QTL was independently validated. Red circles indicate QTL detected in more than one mapping panel.

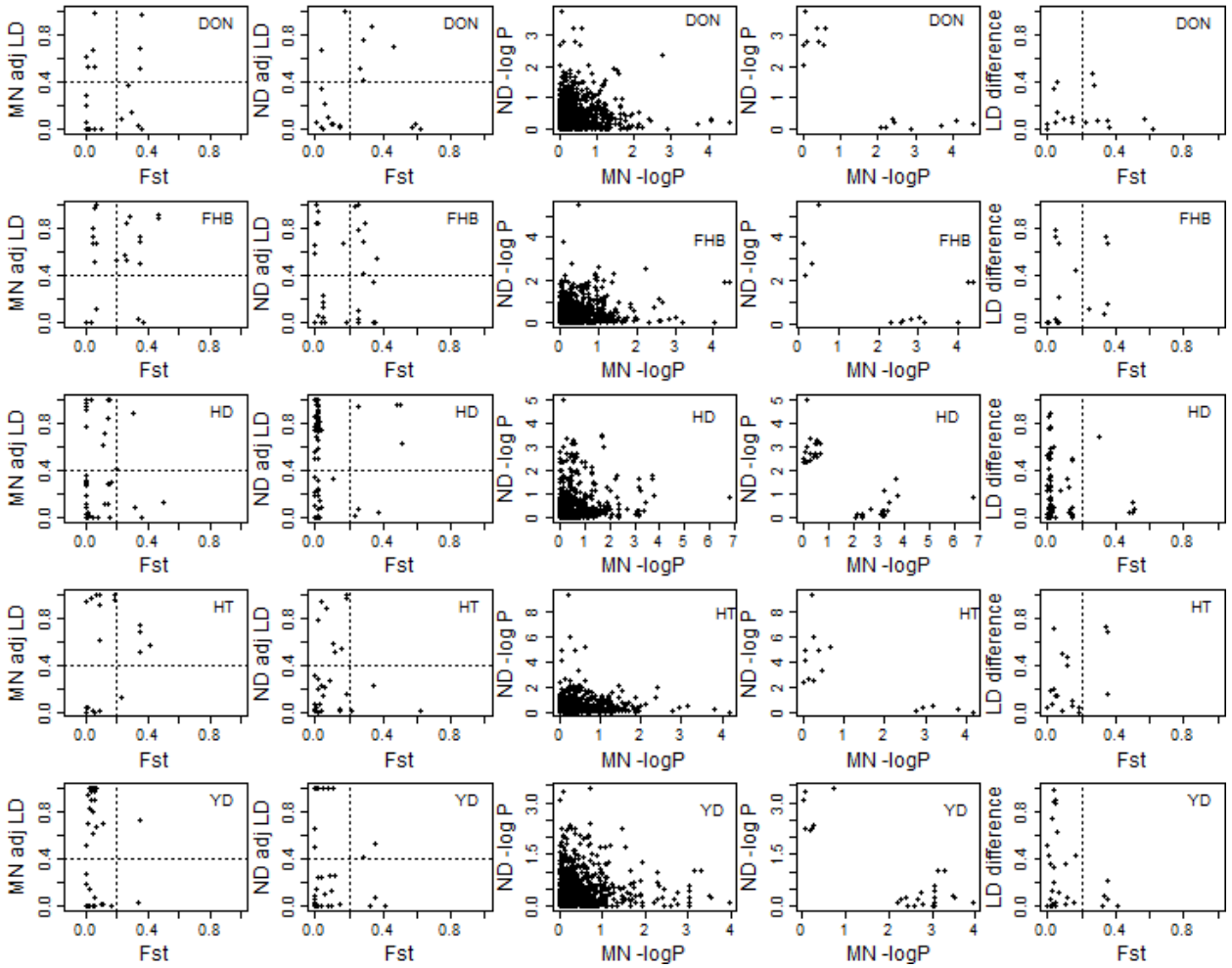

Fig.S8: Relationships among adjacent marker LD, Fst, association mapping significnance value, and LD between the Minnesota and North Dakota breeding programs. Plots are arranged A to E horizontally for each trait.
(A) Details of SNP with higher significance (-logp > 1.75) in MN breeding program. Adjacent marker LD of SNP were plotted against their FSt values calculated using MN and ND breeding programs.
(B) Same as A, but for ND breeding lines.
(C) –logp value of all SNP in MN and ND breeding programs.
(D) Same as C, but only SNP that differed largely (absolute difference in –logp between MN and ND > 2) among breeding programs are shown.
(E) Details of SNP in D. Absolute difference in adjacent marker LD of SNP were plotted against their FSt values calculated using MN and ND breeding programs.

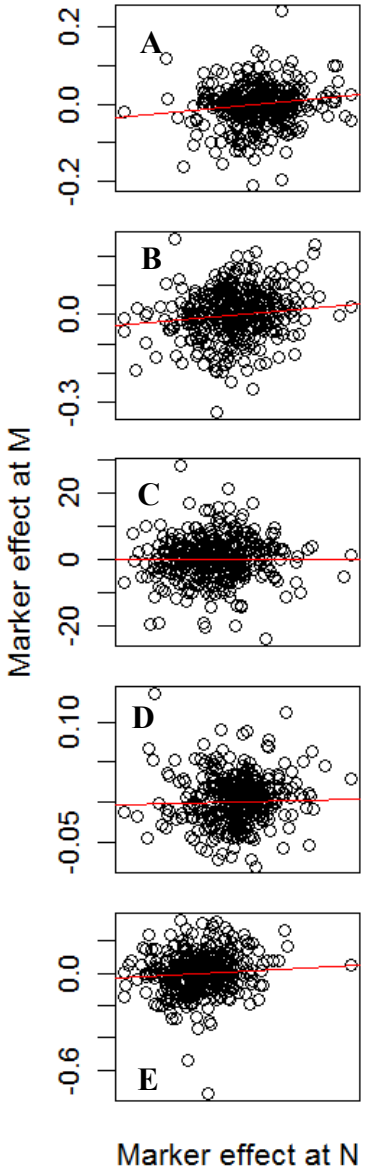

Fig.S9: Pairwise correlation of genomic estimated allelic effects of 300 progeny lines estimated using MN and ND training population. (A) FHB, (B) DON, (C) grain yield, (D) heading date, (E) plant height.

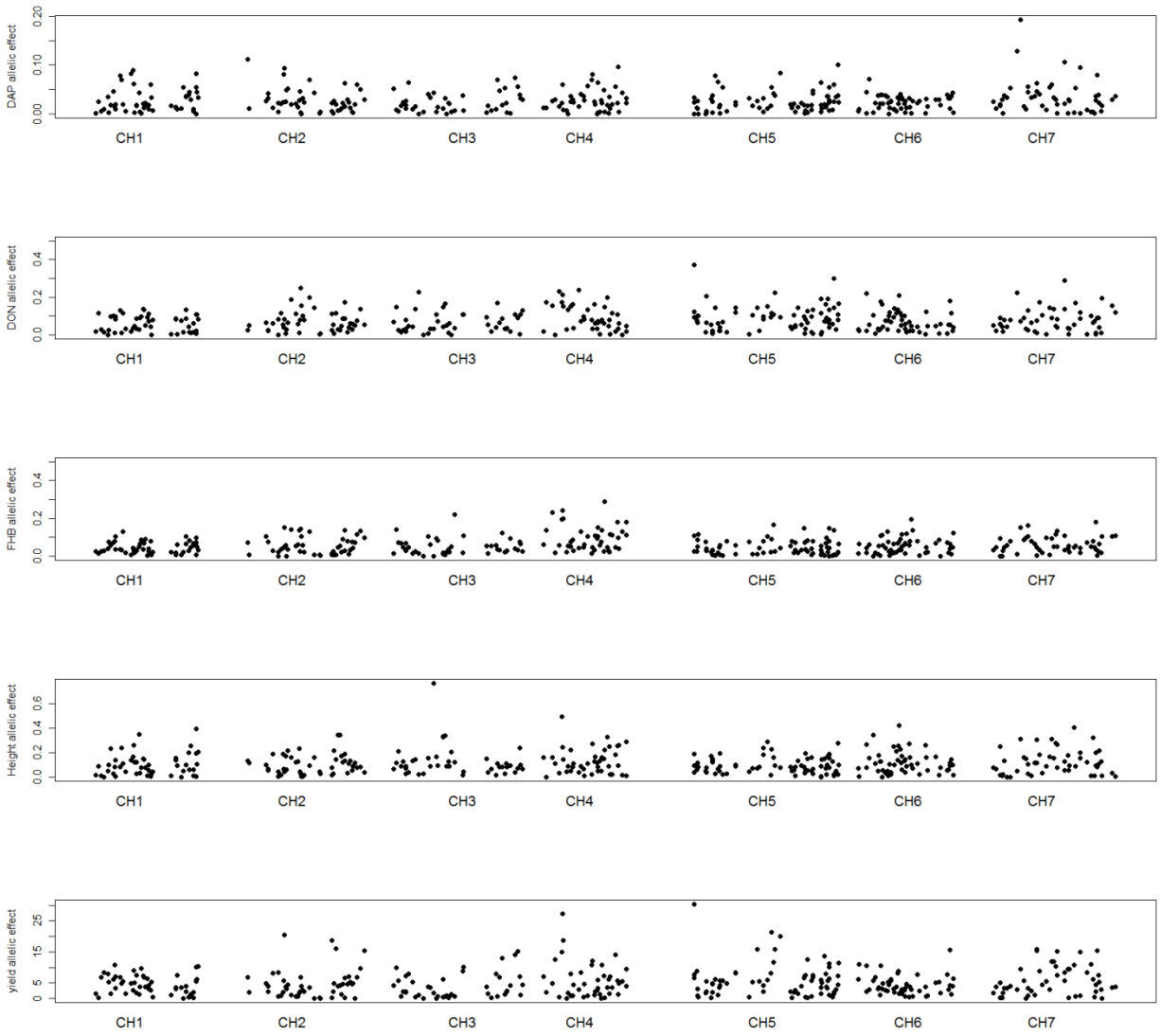

Fig.S10: Absolute difference between allelic effects of markers estimated using MN and ND training populations.

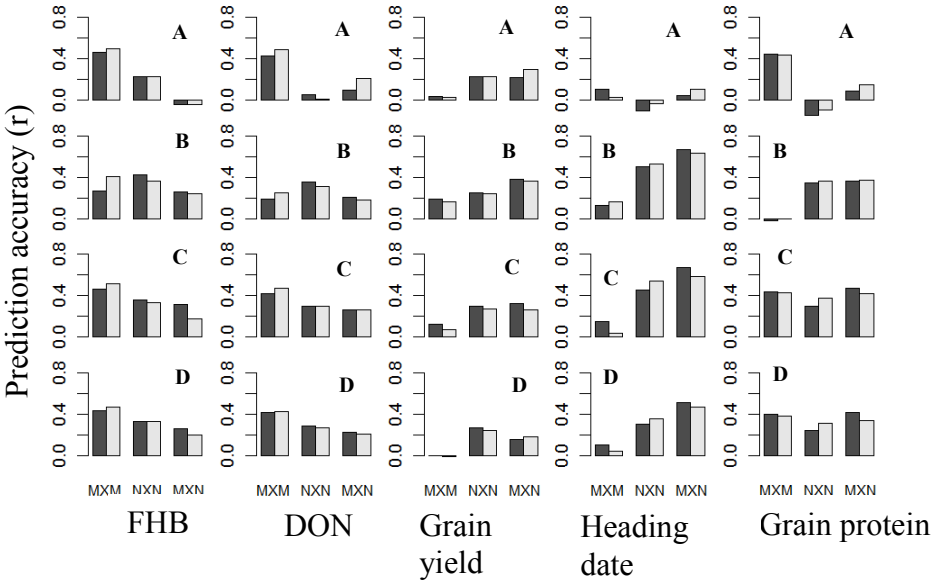

Fig.S11: Comparison of prediction accuracy for three different cross types (MxM, NxN and MxN) using four training populations from the University of Minnesota (A), North Dakota State University (B), both programs combined where n=768 (C), and both programs combined but n=384 (D) . Black bar refers to RR model and grey bar refers to Gauss model.

Table S1 List of barley CAP trials (2006-09) for agronomic traits reported in chapter 1 and chapter 2.

Barley CAP trial name	Trait		
Expt1_2008_Carringt	Heading date	Height	Yield
Expt1_2008_Fargo	HD	Height	Yield
Expt1_2008_Minot	HD	Height	Yield
Expt21_2007_Osnabro	HD	Height	Yield
Expt21_2009_Fargo	HD	Height	Yield
Expt22_2007_Osnabro	HD	Height	Yield
Expt2_2008_Carringt	HD	Height	Yield
Expt2_2008_Fargo	HD	Height	Yield
Expt2_2008_Minot	HD	Height	Yield
Expt2_2008_Ray	HD	Height	Yield
Expt3_2006_Carringt	HD	Height	Yield
Expt3_2006_Fargo	HD	Height	Yield
Expt3_2006_Minot	HD	Height	Yield
Expt3_2006_NessonVa	HD	Height	Yield
Expt3_2006_Osnabroc	HD	Height	Yield
Expt3_2006_Sidney	HD	Height	Yield
Expt3_2006_Willisto	HD	Height	Yield
Expt3_2008_Carringt	HD	Height	Yield
Expt3_2008_Fargo	HD	Height	Yield
Expt3_2008_Minot	HD	Height	Yield
Expt3_2009_Carringt	HD	—	—
Expt3_2009_Fargo	HD	Height	Yield
Expt3_2009_Minot	HD	Height	Yield
Expt3_2009_Ray	HD	Height	Yield
Expt3_2009_Sidney	HD	Height	Yield
Expt3_2009_Willisto	HD	Height	Yield
Expt5_2006_Minot	HD	Height	Yield
Expt5_2006_NessonVa	HD	Height	Yield
Expt5_2006_Sidney	HD	Height	Yield
Expt5_2006_Willisto	HD	Height	Yield
PYT1_2006_Crookston	HD	Height	Yield
PYT1_2006_Morris	HD	Height	Yield
PYT1_2006_StPaul	HD	Height	Yield
PYT1_2007_Crookston	HD	Height	Yield
PYT1_2007_Morris	HD	Height	Yield
PYT1_2007_StPaul	HD	Height	Yield
PYT1_2008_Crookston	HD	Height	Yield
PYT1_2008_StPaul	HD	Height	Yield
PYT1_2009_Crookston	HD	Height	Yield
PYT1_2009_StPaul	HD	Height	Yield

Table S2 List of barley CAP trials (2006-09) for FHB and DON reported in chapter 1 and chapter 2.

Barley CAP trial name	Trait	
2006_StPaulFHB	FHB	—
FHBN_2006_Crookston07	FHB	DON
FHBN_2006_Fargo	FHB	DON
FHBN_2006_Langdon	FHB	DON
FHBN_2007_Crookston	FHB	DON
FHBN_2007_StPaul	FHB	—
FHBN-N6_2007_Osnabrock	FHB	DON
FHBN-MN_2007_Osnabrock	FHB	DON
2008_Morris	FHB	DON
FHBN-MN_2008_Fargo	FHB	DON
FHBN-N6_2008_Fargo	FHB	DON
FHBN_2008_Crookston	FHB	DON
FHBN_2009_Fargo	FHB	DON
FHBN_2009_Langdon	FHB	DON
FHB-MN_2009_Langdon	FHB	DON
FHB-N6-Expt21_2009_Langdo	FHB	DON
FHB-N6-Expt3_2009_Langdon	FHB	DON

Table S3 List of 300 progeny panel test locations (2011 and 2012).

Trial	Trait measured				
ST Paul 2011	DON	FHB	Yield	Heading date	Plant height
ST Paul 2012	—	FHB	Yield	Heading date	Plant height
CROOKSTON 2011	DON	FHB	Yield	Heading date	Plant height
CROOKSTON 2012	—	FHB	Yield	Heading date	Plant height
Nesson valley 2011	—	—	Yield	Heading date	Plant height
Nesson valley 2012	—	—	Yield	Heading date	Plant height
Fargo 2012	—	—	Yield	Heading date	Plant height
Osnabrock 2011	DON	—	—	—	—
Langdon 2011	DON	—	—	—	—
Langdon2012	—	FHB	—	—	—
Osnabrock2012	—	FHB	—	—	—

Table S4: List of SNP markers used to genotype progeny panel. Out of 384 markers used for genotyping, 21 markers failed during genotyping. From the remaining 363 markers listed here, we removed 23 SNPs that are fixed before any analysis. They are: 11_10081, 11_10398, 11_10586, 11_10686, 11_10900, 11_11147, 11_11243, 11_20220, 11_20686, 11_20755, 11_20826, 11_20840, 11_20844, 11_20968, 11_21281, 12_11173, 12_30055, 12_30113, 12_30382, 12_30900, 12_30948, 12_31285, 12_31350.

#	snp	CH	genpos	#	snp	CH	genpos
1	11_10460	1	0.95	46	11_10782	1	131.89
2	12_10636	1	4.51	47	11_10041	1	135.56
3	11_21174	1	8.29	48	12_11496	1	135.56
4	12_30951	1	11.42	49	12_30277	1	135.56
5	12_30948	1	17.26	50	11_20594	1	136.31
6	11_10030	1	18.05	51	11_20840	1	137.83
7	11_20712	1	20.82	52	12_31224	2	4.08
8	11_10186	1	23.86	53	11_10326	2	6.45
9	11_10744	1	26.11	54	11_21265	2	28.44
10	11_21048	1	26.58	55	11_10787	2	31.02
11	12_31276	1	27.35	56	11_20864	2	31.72
12	11_20617	1	33.61	57	11_10525	2	38.03
13	11_10760	1	34.83	58	11_21338	2	44.84
14	12_31177	1	36.95	59	12_30363	2	45.55
15	12_11498	1	40.99	60	12_30703	2	49.03
16	11_10526	1	47.47	61	12_30604	2	51.75
17	12_30043	1	50.6	62	11_10638	2	52.47
18	11_11484	1	51.23	63	12_31474	2	53.53
19	11_20810	1	52.46	64	12_30338	2	54.95
20	11_21217	1	54.73	65	12_11272	2	55.67
21	11_10075	1	55.49	66	12_10485	2	58.24
22	11_21053	1	58.9	67	11_20690	2	62.82
23	11_10552	1	59.71	68	12_10545	2	69.13
24	11_10798	1	61.53	69	11_20833	2	71.12
25	11_20432	1	64.3	70	11_20528	2	73.04
26	11_20642	1	64.91	71	12_10650	2	73.75
27	12_10960	1	66.7	72	11_20734	2	75.18
28	12_30298	1	69.53	73	12_30178	2	75.89
29	11_10686	1	71.43	74	11_10196	2	78.03
30	12_30742	1	71.43	75	12_20489	2	79.19
31	11_21126	1	73.94	76	12_30900	2	86.63
32	11_10279	1	75.45	77	11_10214	2	93.5
33	11_20657	1	77.29	78	11_10398	2	100.37
34	12_11173	1	101.45	79	11_10900	2	101.78
35	11_20220	1	107.55	80	11_20182	2	116.49
36	11_20844	1	108.31	81	12_31095	2	116.49
37	11_20021	1	109.82	82	11_10780	2	119.05
38	12_10905	1	114.84	83	11_21220	2	120.02
39	11_10854	1	117.8	84	12_30636	2	122.21
40	11_21038	1	121.12	85	11_21370	2	125.46
41	11_10586	1	121.77	86	11_10109	2	127.64
42	11_10722	1	125.27	87	11_10065	2	130.01
43	12_31377	1	126.48	88	11_20895	2	131.77
44	11_11038	1	128.14	89	11_20715	2	133.94
45	11_20383	1	131.15	90	12_30396	2	133.94

#	snp	CH	genpos
91	11_20590	2	137.51
92	11_10551	2	139.65
93	12_10447	2	141.28
94	11_21250	2	144.31
95	11_11380	2	145.03
96	11_20293	2	147.94
97	11_21436	2	150.67
98	11_21453	2	155.3
99	12_30102	2	160.29
100	12_31428	3	0
101	11_20159	3	2.9
102	11_20252	3	6.03
103	12_30818	3	9.63
104	11_20595	3	12.46
105	12_10571	3	15.55
106	12_30113	3	15.55
107	11_20172	3	16.33
108	11_20742	3	19.15
109	12_30920	3	24.99
110	11_20968	3	28.44
111	12_30571	3	32.83
112	11_10081	3	39.45
113	12_30064	3	46.31
114	12_30474	3	48.63
115	11_11086	3	53.27
116	11_10926	3	56.4
117	11_11016	3	58.64
118	11_11391	3	65.52
119	11_20704	3	65.52
120	12_30788	3	68.32
121	11_11314	3	70.23
122	12_31356	3	73.53
123	11_20521	3	74.15
124	12_31346	3	76.98
125	12_31262	3	81.66
126	11_10253	3	91.88
127	11_11021	3	93.43
128	12_31499	3	123.68
129	12_30084	3	126.27
130	11_11141	3	130.82
131	12_10122	3	136.66
132	12_11154	3	138.83
133	11_20851	3	141.54
134	11_10631	3	144.64
135	11_21266	3	148.89
136	11_21272	3	150.37
137	12_21376	3	151.22
138	12_30921	3	155.85
139	12_30767	3	162.15
140	11_20605	3	166.22
141	12_30736	3	168.4
142	11_11516	3	169.32
143	12_30055	3	172.41
144	12_30764	4	0.74
145	11_10208	4	3.74
146	12_31458	4	12.02
147	12_30540	4	15.75
148	11_10221	4	21.61
149	11_20422	4	24.59
150	11_21418	4	26.19
151	11_20777	4	26.66
152	12_31313	4	31.43
153	11_21397	4	33.38
154	12_10860	4	36.37
155	12_31524	4	37.12
156	11_20012	4	39.76
157	11_20114	4	40.36
158	11_11405	4	47.6
159	12_30605	4	50.4
160	11_11114	4	54.25
161	11_11244	4	54.98
162	11_20361	4	59.37
163	11_10639	4	65.05
164	12_30755	4	66
165	11_20580	4	68.21
166	11_11513	4	69.51
167	11_10467	4	72.08
168	12_30994	4	72.82
169	12_31362	4	73.57
170	11_10309	4	76.03
171	11_11004	4	77.31
172	11_10523	4	78.77
173	11_20197	4	81.69
174	11_10724	4	82.42
175	11_11213	4	86.27
176	11_20765	4	87.49
177	11_20358	4	88.22
178	11_10588	4	89.39
179	11_20384	4	91.78
180	12_30117	4	96.59

#	snp	CH	genpos
181	11_20119	4	99.28
182	12_30988	4	100.74
183	11_20454	4	101.62
184	11_20974	4	106.03
185	12_31138	4	111.07
186	11_11299	4	111.68
187	11_20226	5	2.09
188	11_20894	5	2.09
189	12_30976	5	2.09
190	12_30979	5	5.68
191	11_20206	5	6.4
192	11_21202	5	7.03
193	11_11381	5	7.48
194	11_20533	5	17.38
195	12_31094	5	18.09
196	11_20010	5	18.72
197	11_10695	5	25.23
198	11_21065	5	26.28
199	11_10974	5	27
200	11_21426	5	27
201	11_11048	5	29.9
202	11_10688	5	34.25
203	11_10580	5	35.69
204	11_10621	5	37.11
205	12_30410	5	37.11
206	11_20845	5	39.97
207	11_20903	5	46.23
208	11_20239	5	57.36
209	12_10079	5	57.98
210	11_20367	5	75.4
211	11_20236	5	80.61
212	11_11355	5	86.63
213	11_21480	5	89.38
214	11_11290	5	94.43
215	11_10578	5	95.08
216	11_20449	5	100.28
217	11_11350	5	104.5
218	11_20134	5	106.16
219	11_20795	5	108.63
220	12_10507	5	110.26
221	11_11200	5	117.47
222	12_30067	5	131.56
223	11_10705	5	132.63
224	11_10783	5	135.72
225	11_21241	5	137.16
226	12_30929	5	137.16
227	12_30635	5	140.76
228	12_30833	5	145.35
229	12_30400	5	149.1
230	11_20388	5	150.34
231	11_20100	5	151.36
232	12_10333	5	151.36
233	11_20104	5	153.51
234	11_11490	5	153.6
235	11_20545	5	159.79
236	12_30759	5	159.79
237	11_20826	5	161.41
238	11_20646	5	161.58
239	12_31375	5	161.58
240	11_11216	5	171.66
241	12_30566	5	171.66
242	11_20686	5	172.38
243	11_10869	5	173.08
244	11_10600	5	176.62
245	12_11450	5	178.43
246	11_21138	5	179.64
247	12_30494	5	180.71
248	11_20189	5	181.43
249	11_20897	5	182.88
250	12_30577	5	182.88
251	12_30769	5	182.88
252	12_31352	5	182.88
253	11_21155	5	187.38
254	11_21052	5	189.6
255	12_31292	5	189.6
256	11_10401	5	191.97
257	12_30382	5	194.64
258	12_10857	5	194.84
259	12_30958	5	196.12
260	11_10669	6	2.27
261	11_21521	6	3.11
262	11_11479	6	12.54
263	11_20415	6	13.21
264	12_10554	6	16.97
265	11_21246	6	22.35
266	11_10136	6	24.36
267	11_10676	6	28.39
268	12_30697	6	29.05
269	12_31308	6	30.72
270	11_10994	6	31.73
271	11_10939	6	33.74
272	11_10427	6	34.4

#	snp	CH	genpos
273	12_30521	6	38.42
274	12_30361	6	40.79
275	11_10129	6	42.36
276	11_20936	6	43.15
277	11_21281	6	43.83
278	12_30317	6	45.44
279	11_10539	6	46.11
280	12_30751	6	49.4
281	12_30782	6	50.07
282	11_10003	6	52.75
283	11_20656	6	54.6
284	12_30511	6	55.94
285	12_30144	6	56.48
286	11_11067	6	58.01
287	11_20266	6	59.56
288	11_10270	6	60.23
289	11_20058	6	60.23
290	11_21069	6	63.95
291	11_20904	6	64.36
292	11_10040	6	65.03
293	11_20714	6	67.04
294	11_20673	6	70.04
295	11_11349	6	71.08
296	12_30940	6	72.54
297	11_20889	6	75.21
298	11_11458	6	81.17
299	11_11147	6	83.89
300	12_31235	6	91.79
301	11_20996	6	93.12
302	11_20972	6	94.73
303	11_20036	6	105.6
304	11_20355	6	110.32
305	11_10239	6	112.32
306	12_31126	6	121.22
307	11_20005	6	122.53
308	11_20868	6	124.85
309	12_30414	6	124.85
310	11_21112	6	126.85
311	11_11111	6	128.48
312	12_30956	6	129.38
313	12_31350	7	3.34
314	12_11433	7	6.78
315	11_20245	7	12.42
316	12_31450	7	12.42
317	12_31285	7	14.96
318	11_20755	7	15.93
319	11_21437	7	17.2
320	11_10025	7	21.13
321	12_30530	7	25.93
322	11_20192	7	34.82
323	12_10218	7	39.04
324	12_10979	7	43.38
325	11_21528	7	46.19
326	11_21491	7	48.9
327	12_30528	7	49.68
328	12_30545	7	53.6
329	12_10959	7	58.57
330	11_11014	7	60.69
331	12_30880	7	61.32
332	12_10605	7	64.8
333	11_11098	7	68.46
334	12_30496	7	73.75
335	12_11477	7	78.22
336	11_10534	7	80.94
337	11_20349	7	83.44
338	12_30199	7	86.44
339	11_10143	7	87.97
340	11_10303	7	87.97
341	11_21201	7	98.5
342	11_21448	7	98.5
343	11_20103	7	102.85
344	12_30630	7	103.62
345	11_10169	7	104.78
346	11_20092	7	110.99
347	12_10241	7	112.46
348	12_30368	7	118.9
349	12_30164	7	119.54
350	11_11243	7	125.55
351	11_21229	7	128.36
352	11_10861	7	133.79
353	11_10078	7	136.62
354	12_30380	7	138.17
355	11_10454	7	140.21
356	12_31282	7	140.21
357	11_21280	7	141.76
358	12_31491	7	143.68
359	11_21223	7	144.45
360	11_10130	7	147.48
361	11_10896	7	148.25
362	11_20170	7	161.54
363	12_30826	7	166.56